

Chen, Xiaohong; Jacho-Chàvez, David T.; Linton, Oliver

Working Paper

Averaging of moment condition estimators

cemmap working paper, No. CWP26/12

Provided in Cooperation with:

The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chen, Xiaohong; Jacho-Chàvez, David T.; Linton, Oliver (2012) : Averaging of moment condition estimators, cemmap working paper, No. CWP26/12, Centre for Microdata Methods and Practice (cemmap), London, <https://doi.org/10.1920/wp.cem.2012.2612>

This Version is available at:

<https://hdl.handle.net/10419/64693>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Averaging of moment condition estimators

Xiaohong Chen
David T. Jacho-Chàvez
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP26/12

Averaging of Moment Condition Estimators*

Xiaohong Chen[†]
Yale University

David T. Jacho-Chávez[‡]
Emory University

Oliver Linton[§]
University of Cambridge

Abstract

We establish the consistency and asymptotic normality for a class of estimators that are linear combinations of a set of \sqrt{n} -consistent estimators whose cardinality increases with sample size. A special case of our framework corresponds to the conditional moment restriction and the implied estimator in that case is shown to achieve the semiparametric efficiency bound. The proofs do not rely on smoothness of underlying criterion functions.

Journal of Economic Literature Classification: C12, C13, C14

Keywords and phrases: Instrumental Variables; Minimum Distance; Semiparametric Efficiency; Two-Stage Least Squares

*Earlier versions of this paper circulated under the title “An Alternative Way of Computing Efficient Instrumental Variable Estimators.” We would like to thank the Co-Editor Guido Kuersteiner, as well as four anonymous referees for valuable comments and suggestions. Juan Carlos Escanciano, Javier Hidalgo, Roger Koenker, Benno Pötscher, Tom Rothenberg and Pravin Trivedi also provided many helpful suggestions. We thank STICERD, the NSF and the ESRC for financial support.

[†]Department of Economics, Yale University, PO Box 208281, New Haven CT 06520-8281, USA. E-mail: xiaohong.chen@yale.edu. Web Page: <http://cowles.econ.yale.edu/faculty/chen.htm>.

[‡]Corresponding author: Department of Economics, Emory University, Rich Building 331, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: djachocho@emory.edu. Web Page: <http://userwww.service.emory.edu/~djachoc/>.

[§]Department of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom, e-mail: obl20@cam.ac.uk. Web Page: <https://sites.google.com/site/oliverlinton/oliver-linton>. This paper was partly written while I was at Universidad Carlos III de Madrid-Banco Santander Chair of Excellence, and I thank them for financial support.

1 Introduction

In this paper we derive the properties of an estimator formed by taking linear combinations of an increasing number of inefficient but \sqrt{n} -consistent estimators obtained from a sequence of moment restrictions. The proposed methodology has the advantage that one can see how much variation there is in the parameter estimates, and how much weight an optimal combination would place on them. In cases where there is truly little variation, the practitioner can presumably do with very simple inference rules. The new estimator is also liable to be useful in high-dimensional models where there is a larger set of instruments than sample observations. A leading example is where the model information is a conditional moment restriction, which generates an infinite number of unconditional moment restrictions. The usual approach here is to combine the unconditional moment restrictions into a single estimating equation; our proposal involves estimating the parameters several times from subsets of the moment conditions and then combining the resulting estimators.

The idea of combining estimates is not new, and has been used to improve finite sample properties of estimators and forecasts. [Granger and Jeon \(2004\)](#) provides an useful discussion. For example, [Sawa \(1973\)](#) considered combining k-class estimators in simultaneous equations systems, for the reason of reducing bias. [Breiman \(1996, 1999\)](#) introduced the idea of bagging, which is based on using bootstrap resamples to compute a large(ish) sample of subsample estimators and then combining them. [Watson \(2003\)](#) and [Stock and Watson \(1999\)](#) propose various methods for combining large numbers of predictors to improve forecasting performance. In the nonparametric literature, [Gray and Schucany \(1972\)](#) and [Bierens \(1987\)](#) have proposed jackknife estimators that combine different kernel smoothers in order to reduce bias. Similarly, [Kotlyarova and Zinde-Walsh \(2006, 2007\)](#) and [Schafgans and Zinde-Walsh \(2010\)](#) have proposed combining kernel smoothers calculated with different bandwidths and kernel functions to construct robust estimators of densities and density-weighted average derivatives respectively.

Our method is in effect a generalization of the classical method of minimum chi-squared or minimum distance discussed in [Malinvaud \(1966\)](#) and [Rothenberg \(1973\)](#), which was conceived as a way of imposing equality restrictions in estimation via first estimating an unrestricted model and then finding the best combination of the unrestricted estimators that imposes the restrictions. In a number of cases this strategy is preferable to solving the constrained estimation problem directly. In our case, the best combination is linear with weights that add up to one.

There is a vast literature on estimating models defined through conditional moment restrictions. We just mention one recent paper that is particularly relevant to our study, [Koenker and Machado](#)

(1999). They considered a similar problem albeit restricted to certain linear models and to a rather specific estimator. They proved that a sufficient condition for the usual asymptotics for generalized method of moments estimation GMM to be valid when the number of unconditional moment equations τ increases with n is that $\tau^3/n \rightarrow 0$.¹ Their results can be interpreted as a warning not to include too many moment conditions in GMM: that the consequences of so doing are not just that no improvement is made, but that the distributional approximation can potentially break down. Our objective is quite different and we deal with nonlinear models.²

In linear models and with efficiency in mind, the proposed method can also be viewed as an alternative to choosing a subset of instruments among a large class of valid instruments, see e.g. Donald and Newey (2001) and Kuersteiner and Okui (2010). For example, consider the case where an unknown but *fixed* number of instruments yields non-identified (Lobato and Dominguez, 2004) or weakly-identified (Stock and Wright, 2000) unconditional moment restrictions, then a simple averaging would make their first-order impact vanish with sample size. On the other hand, if efficiency is not of primary importance, knowledge of the quality of instruments can be readily incorporated into the proposed estimator via the weighting scheme.

We first establish consistency and \sqrt{n} -asymptotic normality of a class of estimators that involve finite linear combinations of an infinite dimensional set of estimators, where the cardinality of the linear combinations increases with sample size. The class of estimators considered is allowed to include those computed from discontinuous criterion functions that are nonlinear in the parameters and data. We also establish that a member of our class of estimators achieves the semiparametric efficiency bound for the conditional moment model. We propose a scheme for estimating the optimal weights and show that this is consistent. We conclude by presenting results of two Monte Carlo experiments showing how our procedure works in practice.

We use $\|A\| = (\text{tr}(A^\top A))^{1/2}$ for any matrix A . Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of a real symmetric matrix A .

¹Also related is work by Donald, Imbens, and Newey (2003) who transform conditional moment restriction into increasing number of unconditional moment equations, and obtain efficiency and consistent asymptotic variance estimation under $\tau^2/n \rightarrow 0$ instead.

²We do not search for the largest value of τ consistent with our asymptotics, although of course the Koenker and Machado (1999) results provide an upper bound.

2 The Model Framework and Estimation

We observe an independent and identically distributed sample $\{Z_i\}_{i=1}^n \in \mathbb{R}^d$. We suppose that there are a sequence of moment conditions g_1, g_2, \dots with $g_j \in \mathbb{R}^q$ such that for some unique $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ we have

$$E[g_j(Z_i, \theta_0)] = 0.$$

For simplicity, we shall assume that $q = p$ so that the j^{th} problem is exactly identified. We define the estimators $\hat{\theta}_j$, $j = 1, 2, \dots$, as any sequence that satisfies

$$G_{nj}(\hat{\theta}_j) = \frac{1}{n} \sum_{i=1}^n g_j(Z_i, \hat{\theta}_j) = o_p(n^{-1/2}). \quad (2.1)$$

For each j , this problem is completely parametric and will result in a \sqrt{n} -consistent and asymptotically normal estimator $\hat{\theta}_j$ (under standard conditions).

We consider a subset $\mathcal{J}_n = \{j_1, \dots, j_\tau\}$ of all possible such estimators, where $\tau = \tau(n)$ is a truncation parameter. We combine these estimators in a linear fashion to produce a new estimator

$$\hat{\theta} = \sum_{j \in \mathcal{J}_n} W_{nj} \hat{\theta}_j, \quad (2.2)$$

where W_{nj} are some matrix weights, possibly stochastic, that sum to the identity. This defines a class of estimators \mathcal{E} indexed by the weighting matrices $\{W_{nj}, j \in \mathcal{J}_n\}$; as we show below, there is a choice of weights that will minimize the asymptotic variance of the estimator $\hat{\theta}$. In some special cases we can show that the resulting estimator will achieve a semiparametric efficiency bound. Although, we later on set $\mathcal{J}_n = \{1, \dots, \tau(n)\}$, the derived asymptotic properties below hold for general sets \mathcal{J}_n .

The estimator (2.2) is a form of minimum distance where the number of restrictions could increase with sample size.³ Even though each criterion function G_{nj} is a nonlinear function of θ , the computational costs of this procedure may not be so great, since one can use the estimates in one step as starting values in the computation of the next step. Additional computational issues arise in connection with the weights W_{nj} but these are discussed below.

There are two tasks we now pursue. The first is to prove that such an estimator (2.2) is consistent and \sqrt{n} -asymptotically normal under general conditions on the truncation parameter and weighting sequence. The second task is to determine the optimal choice of weights. We next consider some examples.

³See [Rothenberg \(1973\)](#) and [Newey and McFadden \(1994\)](#) for finite fixed τ .

3 Examples

Example 1 (Classical two stage least squares in simultaneous equations)

Suppose that⁴

$$y_{1i} = \theta y_{2i} + \varepsilon_i; \quad y_{2i} = \pi_2^\top X_i + u_i,$$

where $(\varepsilon_i, u_i)^\top$ are i.i.d. error terms, $E[\varepsilon_i|X_i] = 0$, $E[u_i|X_i] = 0$ and $X_i \in \mathbb{R}^k$. The two stage least squares estimator is

$$\tilde{\theta} = \frac{\sum_{i=1}^n \hat{y}_{2i} y_{1i}}{\sum_{i=1}^n (\hat{y}_{2i})^2} = \frac{\sum_{i=1}^n \hat{y}_{2i} y_{1i}}{\sum_{i=1}^n \hat{y}_{2i} y_{2i}}, \quad (3.1)$$

where $\hat{y}_{2i} = \hat{\pi}_2^\top X_i$ and $\hat{\pi}_2$ is the vector of least squares estimates obtained from the reduced form regression of y_{2i} on all the instruments $X_i = (X_{1i}, \dots, X_{ki})^\top$. Our estimator is

$$\hat{\theta} = \sum_{j=1}^k W_{nj} \hat{\theta}_j, \quad (3.2)$$

where

$$\hat{\theta}_j = \frac{\sum_{i=1}^n \hat{y}_{2i}^j y_{1i}}{\sum_{i=1}^n (\hat{y}_{2i}^j)^2} = \frac{\sum_{i=1}^n \hat{y}_{2i}^j y_{1i}}{\sum_{i=1}^n \hat{y}_{2i}^j y_{2i}}, \quad (3.3)$$

where $\hat{y}_{2i}^j = \hat{\pi}_{2j}^\top X_{ji}$, and $\hat{\pi}_{2j}$ is the least squares estimates obtained from the reduced form regression of y_{2i} on the single instrument X_{ji} for $j = 1, \dots, k$. Here, W_{nj} are scalar weights that satisfy $\sum_{j=1}^k W_{nj} = 1$. There is a choice of W_{nj} that makes $\hat{\theta}$ asymptotically equivalent to the 2SLS estimator $\tilde{\theta}$. The classical minimum distance estimator (generalized indirect least squares) exploits the relationship between the reduced form coefficients and the structural parameter, i.e., $\pi_{1j}/\pi_{2j} = \theta$, where $\pi_{\ell j} = E(y_{\ell i} X_{ji})/E(X_{ji}^2)$ are the parameters of the reduced form of $y_{\ell i}$ on X_{ji} for $\ell = 1, 2$ and $j = 1, \dots, k$ (the estimator is a linear combination of $\hat{\pi}_{1j}/\hat{\pi}_{2j}$, where $\hat{\pi}_{\ell j}$ are the corresponding reduced form estimators), see [Rothenberg \(1973\)](#).⁵

Example 2 (Infinite order regression model)

Now consider the infinite order least squares regression model

$$Y_i = \sum_{k=1}^{\infty} X_{ki} \beta_k(\theta) + \varepsilon_i, \quad (3.4)$$

⁴For this particular linear model, a closely related paper to ours is [Lee and Zhou \(2011\)](#).

⁵It may be that the moments of $\hat{\theta}_j$ defined in this way do not exist in finite samples, see e.g., [Phillips \(1983\)](#). To avoid this issue, one could divide the instruments into groups with two or more members, estimate the individual 2SLS within the group, and then average as before.

where θ is some finite dimensional parameter and ε_i is an error term satisfying $E(\varepsilon_i X_{ji}) = 0$, $j = 1, 2, \dots$. Consider the special case that $\beta_k(\theta) = \theta$ for all k . Then, we need at least that $E[(\sum_{k=1}^{\infty} X_{ki})^2] < \infty$ in order for the summation in (3.4) to be well defined; this would be satisfied if $\sigma_k^2 = E(X_{ki})^2$ goes to zero at a rate faster than k^{-1} as $k \rightarrow \infty$. The optimal estimator under homoskedasticity is the OLS estimator of Y_i on $\sum_{k=1}^{\infty} X_{ki}$. If also the regressors are mutually orthogonal, i.e., $E(X_{ji} X_{ki}) = 0$ for all $j \neq k$, the OLS estimators of Y_i on X_{ki} are consistent, and so will any linear combination thereof, and so we can construct estimators of θ by taking linear combinations of these marginal OLS regressions.⁶

3.1 Semiparametric Instrumental Variables

We suppose that $Z_i^\top = (Y_i^\top, X_i^\top)$ and that there is a unique $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ satisfying the conditional moment conditions

$$E[\rho(Z_i, \theta_0) | X_i] = 0 \tag{3.5}$$

with probability one, where $\rho(z, \theta)$ is a scalar residual function. This implies the unconditional moment conditions

$$E[A(X_i)\rho(Z_i, \theta_0)] = 0, \tag{3.6}$$

for any $p \times 1$ measurable vector $A(X_i)$ (for which the expectation exists). The sample version of (3.6) is the basis of estimation as described in many previous papers, including Amemiya (1974) and Hansen (1982).

Suppose that $E[\rho(Z_i, \theta_0)^2 | X_i] = \sigma_0^2(X_i)$ is positive with probability one, and that

$$D_0(X_i) = \left(\frac{\partial}{\partial \theta} E[\rho(Z_i, \theta) | X_i] \right)_{\theta=\theta_0}$$

exists with probability one. In this case, the optimal (instrumental variables) matrix is proportional to $A_{\text{oiv}}(X_i) = D_0(X_i)\sigma_0^{-2}(X_i)$, and the resulting optimal instrumental variables (oiv) –or optimal GMM– estimator $\tilde{\theta}_{\text{oiv}}$ has asymptotic variance $\Sigma_{\text{oiv}} = \{E[\sigma_0^{-2}(X_i)D_0(X_i)D_0(X_i)^\top]\}^{-1}$ - see for example Hansen (1985), Chamberlain (1987), Newey (1990, 1993) for smooth ρ and Chen and Pouzo (2009) for non-smooth ρ .

⁶By changing variables to X_{ki}/σ_k the parameters become $\theta \cdot \sigma_k$ in which case the problem is more like the instrumental variables regression because the regressors have the same variance but the parameters decline in importance.

Suppose that the optimal matrix $A_{\text{oiv}}(\cdot)$ is of unknown form, but can be represented, in an L_2 sense, by the following series expansion

$$A_{\text{oiv}}(x) = D_0(x)\sigma_0^{-2}(x) = \sum_{j=1}^{\infty} \beta_{j0}\phi_j(x),$$

where $\phi_j(\cdot)$ are known basis functions chosen by the practitioner, while β_{j0} are unknown coefficients determined uniquely by the basis. For notational convenience we shall allow ϕ_j to be $p \times 1$ vectors; in general, β_{j0} depends on θ_0 and is a $p \times p$ matrix. A common approach here is to estimate the coefficients β_{j0} (by say series approximation, see e.g. [Newey, 1990](#)) and then to let

$$\widehat{A}_\theta(x) = \sum_{j=1}^{\tau(n)} \widehat{\beta}_j(\theta)\phi_j(x),$$

where $\tau(n)$ is some truncation sequence that goes to infinity with sample size but at a slow rate. Then let $\widetilde{\theta}_{\text{oiv}}$ be any sequence that satisfies

$$\frac{1}{n} \sum_{i=1}^n \widehat{A}_{\widetilde{\theta}_{\text{oiv}}}(X_i)\rho(Z_i, \widetilde{\theta}_{\text{oiv}}) = o_p(n^{-1/2}).$$

In current parlance this would be called a continuously updated oiv estimator. An alternative method is to use some preliminary consistent estimator of θ_0 to first construct a consistent estimator of A_{oiv} , and then to solve a similar first order condition with the estimated instrument. [Newey \(1990, 1993\)](#) showed that such an estimator is asymptotically equivalent to the instrumental variable procedure based on knowing the optimal instrument function A_{oiv} and computing solutions $\widetilde{\theta}_{\text{oiv}}$ to

$$\frac{1}{n} \sum_{i=1}^n A_{\text{oiv}}(X_i)\rho(Z_i, \widetilde{\theta}_{\text{oiv}}) = o_p(n^{-1/2}).$$

See [Newey and McFadden \(1994\)](#) for discussion. There have been a number of alternative suggestions made more recently with a view to improving small sample performance, [Newey and Smith \(2004\)](#) contains an excellent review of this literature.

We take a different approach. Instead of estimating the optimal instrument function A_{oiv} we will estimate the optimal way to combine all the estimators defined through the individual moment restrictions. We consider a sequence of pre-specified generic basis ($p \times 1$ vector-valued) functions $\{A_j(\cdot)\}$ ($\phi_j(\cdot)$) such that $E[||A_j(X_i)||^2] < \infty$; for instance, we may take a uniformly bounded basis such as the B-spline basis. Then we solve

$$\frac{1}{n} \sum_{i=1}^n A_j(X_i) \rho(Z_i, \hat{\theta}_j) = o_p(n^{-1/2})$$

for each j and combine as in (2.2).

4 Large Sample Properties

We begin by defining the sample and population first order conditions. For $j = 1, 2, \dots$, let

$$G_{nj}(\theta) \equiv \frac{1}{n} \sum_{i=1}^n g_j(Z_i, \theta) \text{ and } G_j(\theta) \equiv EG_{nj}(\theta).$$

We do not assume that the function $G_{nj}(\theta)$ is differentiable or even continuous, although smoothness conditions are imposed on the expectation $G_j(\theta)$. In this way, we allow also quantile regression estimators (e.g., [Koenker and Bassett, 1978](#)), [Huber's \(1967\)](#) M-estimators, and simulation-based estimators (e.g., [McFadden, 1989](#); [Pakes and Pollard, 1989](#)). For some of the arguments we only require high level conditions on the sample and population first order conditions, and so our results can apply more generally to any linear combination of estimators that have appropriate expansions.

4.1 Consistency

In this subsection we give our consistency result for the estimator (2.2). We make the following assumptions.

ASSUMPTION A: Let $\theta_0 \in \Theta$ satisfy model (3.5).

(A1) The triangular array $\{W_{nj}\}_{j \in \mathcal{J}_n}$, $n = 1, \dots$, satisfies

$$\sum_{j \in \mathcal{J}_n} W_{nj} = I_p \text{ and } \sup_n \sum_{j \in \mathcal{J}_n} \|W_{nj}\| < \infty \text{ w.p.1.} \quad (4.1)$$

Here, $\tau(n)$ satisfies $\tau(n) \rightarrow \infty$ as $n \rightarrow \infty$.

(A2) For all $\delta > 0$ and $n \geq 1$, there is an $\epsilon_n(\delta) > 0$ (with $\epsilon_n(\delta) \rightarrow 0$ as $n \rightarrow \infty$) such that

$$\min_{j \in \mathcal{J}_n} \inf_{\|\theta - \theta_0\| > \delta} \|G_j(\theta)\| \geq \epsilon_n(\delta) > 0.$$

(A3) For the sequences $\epsilon_n(\delta)$, $\tau(n)$ defined above, there exists a positive sequence $\alpha_{1n} = o(1)$ with $\sup_n(\alpha_{1n}/\epsilon_n(\delta)) < \infty$ such that

$$\max_{j \in \mathcal{J}_n} \left(\|G_{nj}(\hat{\theta}_j)\| - \inf_{\theta \in \Theta} \|G_{nj}(\theta)\| \right) = o_p(\alpha_{1n}),$$

(A4) For the sequences $\epsilon_n(\delta)$, $\tau(n)$ defined above, there exists a positive sequence $\alpha_{2n} = o(1)$ with $\sup_n(\alpha_{2n}/\epsilon_n(\delta)) < \infty$ such that

$$\max_{j \in \mathcal{J}_n} \sup_{\theta \in \Theta} \|G_{nj}(\theta) - G_j(\theta)\| = o_p(\alpha_{2n}).$$

The assumptions on the weights are quite weak and are satisfied by many suitable weighting sequences both random and non-random. For example, equal weighting $W_{nj} = 1/\tau(n)I_p$ satisfies the assumption **A1**, where I_p represents a $p \times p$ identity matrix. There are no explicit conditions on the truncation sequence $\tau(n)$ here, but the Assumptions **A2–A4** may require some restrictions on the rate at which $\tau(n)$ increases with n . Assumption **A3** is just a definition of the estimator uniformly over j . Assumption **A2** guarantees that identification is not lost when going from conditional to unconditional moment restrictions, see [Lobato and Dominguez \(2004\)](#) for example.⁷ The identification Assumption **A2** takes account of the fact that each additional moment condition is adding less and less information. The rate at which $\epsilon_n(\delta)$ declines is determined by the sequence $\tau(n)$ and, in the IV case, by the sequence A_j , in particular the rate at which $\|E[A_j(X)]\|$ decreases. By choosing $\tau(n)$ to grow very slowly we can compensate for a rapid decline in the moments of the instruments.

Our conditions require that each member $\hat{\theta}_j$ of the class indexed by \mathcal{J}_n be consistent. This is, however, not strictly necessary by the following arguments. Consider the scalar parameter case where W_{nj} are also scalars and suppose that the parameter space Θ is compact with Euclidean diameter Δ . Let $\mathcal{J}_n = \mathcal{J}_n^c \cup \mathcal{J}_n^I$, where \mathcal{J}_n^c contains $\tau^c(n)$ consistent estimators and \mathcal{J}_n^I contains $\tau^I(n)$ possibly inconsistent estimators. Then, by the triangle inequality

$$|\hat{\theta} - \theta_0| \leq \Delta \sum_{j \in \mathcal{J}_n^I} |W_{nj}| + \sum_{j \in \mathcal{J}_n^c} |W_{nj}| \max_{j \in \mathcal{J}_n^c} |\hat{\theta}_j - \theta_0|.$$

Therefore, it suffices that (4.1) holds, that $\hat{\theta}_j$ are uniformly consistent over the class \mathcal{J}_n^c , and that $\sum_{j \in \mathcal{J}_n^I} |W_{nj}| \rightarrow 0$. Under equal weighting for example, this latter condition would be implied by $\tau^I(n)/\tau^c(n) \rightarrow 0$.

⁷It is a testable restriction, see e.g. [Inoue and Rossi \(2011\)](#) and [Bravo, Escanciano, and Otsu \(2011\)](#).

The uniform convergence Assumption [A4](#) is easy to verify, although it requires one to check the $\max_{j \in \mathcal{J}_n}$ factor. This factor costs little extra, as can be verified from the Bonferroni and exponential inequalities (see below). Since we must have $\epsilon_n(\delta)$ of larger order than $n^{-1/2}$ in the case of i.i.d. data this puts an upper limit on the rate at which $\tau(n)$ can grow, but no lower limit. If $\tau(n)$ only increases very slowly, say like $\log n$, the stated rate is easy to achieve.

Theorem 1 (i) *Suppose that Assumptions [A1–A4](#) hold. Then $\widehat{\theta} - \theta_0 = o_p(1)$.*

For the purpose of obtaining \sqrt{n} -asymptotic normality of $\widehat{\theta}$ in the next subsection, we need to first establish that $\widehat{\theta} - \theta_0 = o_p(n^{-1/4})$ under the following stronger version of Assumption A:

ASSUMPTION A*: Let $\theta_0 \in \Theta$ satisfy model [\(3.5\)](#).

(A*1) [A1](#) holds.

(A*2) There is a positive sequence $\{\gamma_j, j \in \mathcal{J}_n\}$ such that for some $\delta > 0$ and all θ such that $\|\theta - \theta_0\| < \delta$ and $\theta \in \Theta$:

$$\|G_j(\theta)\| \geq \gamma_j \|\theta - \theta_0\|$$

and $\min_{j \in \mathcal{J}_n} \gamma_j \geq \epsilon_n > 0$.

(A*3) For all $\delta_n = o(1)$ and $n \geq 1$,

$$\max_{j \in \mathcal{J}_n} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\|\theta - \theta_0\| \leq \delta_n} \|G_{nj}(\theta)\| \right) = o_p(\epsilon_n n^{-1/4}).$$

(A*4) For all $\delta_n = o(1)$ and $n \geq 1$,

$$\max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n} \|G_{nj}(\theta) - G_j(\theta)\| = o_p(\epsilon_n n^{-1/4}).$$

Assumption [A*2](#) is standard as in [Pakes and Pollard \(1989\)](#), except that we require the lower bounds to decay at a rate under our control. The sequence ϵ_n depends on the sequence of moment conditions but also on the set \mathcal{J}_n : If this set contains few elements then it is possible to make ϵ_n decay very slowly. Assumption [A*3](#) again defines the estimator, while Assumption [A*4](#) requires a rate for the resulting random variable to converge to zero. Both assumptions are similar to those often found in the estimation literature with non-smooth objective functions, see [Newey and McFadden \(1994, Section 7\)](#), with the exception that we are taking a maximum over an increasing number of first order conditions. However, it is likely to be satisfied in most problems. The uniformity across θ is

usually satisfied, indeed we can expect in many cases that $\sup_{\theta \in \Theta} \|G_{nj}(\theta) - G_j(\theta)\| = O_p(1/\sqrt{n})$ for any compact parameter set Θ . Below we provide a Lemma that can be used to verify the uniformity across j condition and may be useful elsewhere.

Theorem 1 (ii) *Suppose that Assumptions [A*1](#)–[A*4](#) hold. Then $\widehat{\theta} - \theta_0 = o_p(n^{-1/4})$.*

Of course there are many alternative ways to impose sufficient conditions which lead to convergence rate. We conclude this subsection with a result that is needed in verifying Assumption [A*4](#) above.

Lemma 1 *Let U_{ji} be a triangular array of random variables, $i = 1, \dots, n$, $j = 1, \dots, \tau(n)$, i.i.d. across i for each j with $E(U_{ji}) = 0$ and $E[|U_{ji}|^\kappa] = c_j < \infty$ for some $\kappa \geq 2$. Let $s_{nj}^2 = \sum_{i=1}^n \text{var}(U_{ji}) = n\sigma_j^2$, where $\sigma_j^2 \rightarrow \infty$ as $j \rightarrow \infty$, and let*

$$a_n = \left(\max_{j \in \mathcal{J}_n} \sigma_j^2\right) \log \tau(n) + \left(\sum_{j \in \mathcal{J}_n} \frac{c_j^2}{\sigma_j^{2\kappa}}\right)^{1/\kappa}. \quad (4.2)$$

Then we have for $\delta_n = a_n \varrho_n$ for any increasing sequence ϱ_n that

$$\max_{j \in \mathcal{J}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n U_{ji} \right| = o_p(\delta_n).$$

For example, if we take $\kappa = 2$, then $a_n = (\max_{j \in \mathcal{J}_n} \sigma_j^2) \log \tau(n) + \sqrt{\tau(n)}$. One application of this Lemma is when $n^{-1/2} \sum_{i=1}^n U_{ji}$ is the leading term of the estimator $\widehat{\theta}_j$, in which case, σ_j^2 would be Γ_j^{-1} (under homoskedasticity) as defined in [\(4.3\)](#) below. Therefore, the corresponding a_n is of order $\Gamma_{\tau(n)}^{-1} \log \tau(n) + \sqrt{\tau(n)}$. Provided $\tau(n)$ does not increase too rapidly, this is less than $n^{1/4}$ as would be required by assumption [A*4](#). Furthermore, it implies that $\max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\|$ goes to zero no slower in probability than $(\Gamma_{\tau(n)}^{-1} \log \tau(n) + \sqrt{\tau(n)})/\sqrt{n}$.

4.2 Asymptotic Normality

In this subsection we derive the asymptotic distribution of our estimator $\widehat{\theta}$, under additional conditions. We strengthen the conditions of [Pakes and Pollard \(1989\)](#) and [Newey and McFadden \(1994\)](#) to accommodate our more general set-up, but again we do not require smoothness conditions on the moment conditions $g_j(Z_i, \theta)$. Then $G_{nj}(\theta) = n^{-1} \sum_{i=1}^n g_j(Z_i, \theta)$ and $G_j(\theta) = E[g_j(Z_i, \theta)]$. We denote

$$\Gamma_j = \frac{\partial}{\partial \theta^\top} G_j(\theta_0) = \frac{\partial}{\partial \theta^\top} E[g_j(Z_i, \theta)] \Big|_{\theta=\theta_0}. \quad (4.3)$$

In the IV example If $D_0(X_i) = \{\partial E[\rho(Z_i, \theta)|X_i]/\partial \theta\}|_{\theta=\theta_0}$ exists with probability one, then we have $\Gamma_j = E[A_j(X_i)D_0(X_i)^\top]$.

ASSUMPTION B: Let $\theta_0 \in \Theta$ satisfy model (3.5).

(B1) $\max_{j \in \mathcal{J}_n} (\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\|\theta - \theta_0\| \leq \delta_n} \|G_{nj}(\theta)\|) = o_p(1/\sqrt{n})$ for any $\delta_n = o(n^{-1/4})$.

(B2) There exists a finite constant C such that for any θ within a shrinking $(n^{-1/4-})$ neighborhood of θ_0

$$\max_{j \in \mathcal{J}_n} \|G_j(\theta) - \Gamma_j(\theta - \theta_0)\| \leq C\|\theta - \theta_0\|^2,$$

where Γ_j is of full (column) rank for each j .

(B3) (a) $\max_{j \in \mathcal{J}_n} \|\sqrt{n}[G_{nj}(\theta_0) - G_j(\theta_0)]\| = O_p(1)$.

(b) For any $\delta_n = o(n^{-1/4})$,

$$\max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n} \|[G_{nj}(\theta) - G_j(\theta)] - [G_{nj}(\theta_0) - G_j(\theta_0)]\| = o_p(1/\sqrt{n}).$$

(B4) There exists a deterministic sequence of matrices W_{nj}^0 satisfying: (a) $\sum_{j \in \mathcal{J}_n} \|(W_{nj} - W_{nj}^0)\Gamma_j^{-1}\| = o_p(1)$; (b) $\limsup_n \sum_{j \in \mathcal{J}_n} \|W_{nj}^0 \Gamma_j^{-1}\| < \infty$.

(B5) (a) The matrix $\Sigma_n = \sum_{j \in \mathcal{J}_n} \sum_{l \in \mathcal{J}_n} W_{nj}^0 V_{jl} W_{nl}^{0\top}$ has a finite positive definite limit Σ , where for all $j, l \in \mathcal{J}_n$,

$$V_{jl} = \Gamma_j^{-1} E[g_j(Z_i, \theta_0) g_l(Z_i, \theta_0)^\top] \Gamma_l^{-1\top}$$

(b) The triangular array of random variables $f_n(Z_i) = n^{-1/2} \sum_{j \in \mathcal{J}_n} c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0)$ satisfies $nE|f_n(Z_i)|^{2+\kappa} \rightarrow 0$ for $\forall c \in \mathbb{R}^p$ and some $\kappa > 0$.

(B6) θ_0 is in the interior of Θ .

(B7) $\max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| = o_p(n^{-1/4})$.

Assumption B1 is again just the definition of the estimator. Assumption B2 requires essentially two uniformly continuous derivatives for the population moment function at $\theta = \theta_0$ and that the first derivative matrix be of full rank.

For Assumption B3(b), consider the empirical distribution function as an example, which satisfies

$$\sup_{|x-x_0| \leq a/n^\alpha} \left| \sqrt{n}[F_n(x) - F(x)] - \sqrt{n}[F_n(x_0) - F(x_0)] \right| = O_p(n^{-\alpha/2})$$

for any $\alpha < 1$ and constant a .⁸ The cost of the additional max is typically no more than an additional factor of order $\sqrt{\tau(n)}$ as is evidenced in Lemma 1 above.

In **B4**, we require that if the weights are random that they can be well approximated by some nonrandom sequence with certain summability properties. This condition entails some restrictions on the rate of growth of τ , and these restrictions can be as much as requiring that $\tau^3/n \rightarrow 0$, see [Koenker and Machado \(1999\)](#). The restrictions are not so stringent in special cases and really arise out of the nonlinearity of the estimating equation rather combined with the large number of parameters.

Assumption **B5** allows us to apply the Liapunov's central limit theorem for triangular arrays to the leading term. This condition is satisfied for a variety of problems, and it implicitly imposes restrictions on how fast $\tau(n)$ could grow with sample size n . Notice that Assumption **B5(b)** is simply: for some $\kappa > 0$ and for all c ,

$$E \left(\left| \sum_{j \in \mathcal{J}_n} c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0) \right|^{2+\kappa} \right) = o(n^{\kappa/2}).$$

For example, suppose we only require that $g_j(Z_i, \theta_0)$ have uniformly bounded fourth moments. Define the positive sequence

$$\epsilon_n = \min_{j \in \mathcal{J}_n} \lambda_{\min}(\Gamma_j).$$

Then, by the Cauchy-Schwarz inequality

$$nE[f_n(Z_i)^4] = \frac{1}{n} \sum_{j,k,l,m \in \mathcal{J}_n} E[\varphi_{ji}\varphi_{ki}\varphi_{li}\varphi_{mi}] \leq \frac{1}{n\epsilon_n^4} \left(\sup_n \sum_{j \in \mathcal{J}_n} \|W_{nj}^0\| \right)^4,$$

where $\varphi_{ji} = c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0)$. It suffices in this case that $n\epsilon_n^4 \rightarrow \infty$. Now suppose that in fact, the scalar $g_j(Z_i, \theta_0)$ are normally distributed with mean zero and variance Γ_j and mutually independent, and that the weights are equal, i.e., $W_{nj}^0 = 1/\tau(n)I_p$ for each j . Then

$$nE[f_n(Z_i)^4] = \frac{1}{n\tau^4} \left(\sum_{j \in \mathcal{J}_n} 3\Gamma_j^{-2} + 3 \sum_{j \neq k \in \mathcal{J}_n} \Gamma_j^{-1}\Gamma_k^{-1} \right) \leq \frac{3}{n\tau^2\epsilon_n^2},$$

which goes to zero provided $n\tau^2\epsilon_n^2 \rightarrow \infty$. These conditions can be weakened considerably in special cases.

⁸We are grateful to Benedikt Pötscher for pointing this out to us. This is due to the Hölder continuity of the limiting Brownian bridge process $B(\cdot)$ of $\sqrt{n}[F_n(\cdot) - F(\cdot)]$, i.e., $|B(x) - B(x_0)| \leq c \cdot |x - x_0|^{1/2}$ for some random variable c with bounded moment. The local uniformity (across i) comes at very little extra cost.

Notice that we can replace Assumptions **B3(a)** and **B5** by the condition that $\{G_{nj}(\theta_0) - G_j(\theta_0) : j \in \mathcal{J}_n\}$ is a Donsker class, i.e., it satisfies the uniform central limit theorem. This kind of assumption has been used in [Portnoy \(1984\)](#) for example.

The condition **B7** that $\max_{j \in \mathcal{J}_n} \|\hat{\theta}_j - \theta_0\| = o_p(n^{-1/4})$ follows from our [Theorem 1\(ii\)](#). It may be possible to prove our result below without a sup-norm convergence result like this, although we have not been able to find a proof based on other convergences like L_p . The usual proofs in other semiparametric estimation problems typically make use of similar results about the convergence of nuisance parameters.

Theorem 2 *Suppose that Assumptions A1 and B1–B7 hold. Then $\sqrt{n}(\hat{\theta} - \theta_0) \implies N(0, \Sigma)$.*

The asymptotic variance matrix Σ depends on the weighting scheme and on the class of estimators considered and, of course, on the underlying distribution of the data. We discuss the nature of the asymptotic variance more in the next section.

To construct consistent estimates of Σ , we would compute

$$\begin{aligned} \hat{\Sigma} &= \sum_{j \in \mathcal{J}_n} \sum_{l \in \mathcal{J}_n} W_{nj} \hat{V}_{jl} W_{nl}^\top \\ \hat{V}_{jl} &= \hat{\Gamma}_j^{-1} \frac{1}{n} \sum_{i=1}^n g_j(Z_i, \hat{\theta}) g_l(Z_i, \hat{\theta})^\top \hat{\Gamma}_l^{-1\top}. \end{aligned} \quad (4.4)$$

Note that there is no further need of a bandwidth parameter here since the cardinality of \mathcal{J}_n is small compared with n . The estimation of Γ_j is easy when G_{nj} are differentiable. In this case,

$$\hat{\Gamma}_j = \frac{1}{n} \sum_{i=1}^n \frac{\partial g_j(Z_i, \hat{\theta})}{\partial \theta} \rightarrow^p \Gamma_j \quad (4.5)$$

under some mild regularity conditions. When G_{nj} are not differentiable, as for example in the Least Absolute Deviation (LAD) case, this method is not feasible. In some cases, one might be able to estimate directly the quantity Γ_j . For example, in the LAD case (with errors independent of covariates), Γ_j is proportional to the density of the errors evaluated at their median. This quantity can be estimated by a variety of nonparametric methods. A general strategy for estimating Γ_j is to use ‘numerical derivatives’, that is, let

$$\hat{\Gamma}_{j;lk} = \frac{1}{n} \sum_{i=1}^n \frac{g_{jl}(Z_i, \hat{\theta} + \delta e_k) - g_{jl}(Z_i, \hat{\theta})}{\delta}, \quad (4.6)$$

where e_k is a vector of zeros with one in the k^{th} position, while δ is a small constant. If we let $\delta(n)$ go to zero at a certain rate as sample size increases, we can show that $\widehat{\Gamma}_{j;lk} \rightarrow^p \Gamma_{j;lk}$, see for example [Pakes and Pollard \(1989\)](#). The actual derivative (4.5) makes δ go to zero before n , but our modified estimator (4.6) allows δ to go to zero with n and indeed slower than n . Under stronger conditions, we can obtain $\max_{j \in \mathcal{J}_n} \|\widehat{\Gamma}_j - \Gamma_j\| \rightarrow^p 0$, $\max_{j,l \in \mathcal{J}_n} \|\widehat{V}_{jl} - V_{jl}\| \rightarrow^p 0$, and $\widehat{\Sigma} \rightarrow^p \Sigma$ to formally justify this procedure. Provided that $\tau(n) \rightarrow \infty$ slowly, the additional conditions are not particularly onerous.

Example 2 (*cont.*)

Suppose that the errors are homoskedastic and the regressors are mutually orthogonal with $E(X_{ji}^2) = \sigma_j^2$. A necessary and sufficient condition for the \sqrt{n} -rate of convergence is that

$$\limsup_{n \rightarrow \infty} \sum_{j \in \mathcal{J}_n} W_{nj}^2 \sigma_j^{-2} < \infty$$

with probability one. Since we also require $\sum_{j=1}^{\infty} \sigma_j^2 < \infty$, this rules out the equal weighting case. Nevertheless, a variety of weighting conditions satisfy the requirement. Furthermore, there is no explicit restriction on τ itself in this case.

5 Optimal Weights

We now discuss the optimal weights in the sense of minimizing asymptotic variance. For simplicity, we restrict attention to the leading case to the case where $\mathcal{J}_n = \{1, \dots, \tau\}$. We first consider the case where τ is fixed and then turn to the case where it is increasing.

5.1 Case 1: Fixed τ

We can consider the optimal weights to be those that minimize the asymptotic variance matrix of Theorem 2 in the special case where τ is fixed, i.e., minimize

$$\Sigma^\tau = \sum_{j=1}^{\tau} \sum_{l=1}^{\tau} W_{nj} V_{jl} W_{nl}^\top$$

with respect to $p \times p$ matrices $W_{n1}, \dots, W_{n\tau}$ subject to the restriction that $\sum_{j=1}^{\tau} W_{nj} = I_p$.⁹ The solution to this can be found explicitly, thus, writing $(B_1, \dots, B_\tau) = (I_p \otimes i_\tau)^\top V^{-1}$, where $V = (V_{j,l})$

⁹This optimization problem can be seen as a multivariate version of the classical portfolio minimum variance choice, which has an explicit solution, see [Campbell, Lo, and Mackinlay \(1997, p. 184-185\)](#).

and i_τ is a τ by 1 vector of ones, we have

$$W_{0j}^{\text{opt}} = \left(\sum_{l=1}^{\tau} B_l \right)^{-1} B_j = \left((I_p \otimes i_\tau)^\top V^{-1} (I_p \otimes i_\tau) \right)^{-1} \left((I_p \otimes i_\tau)^\top V^{-1} \right)_j. \quad (5.1)$$

This results in the asymptotic (as $n \rightarrow \infty$ and τ fixed) variance

$$\Sigma_{\text{opt}}^\tau = \sum_{j=1}^{\tau} \sum_{l=1}^{\tau} W_{0j}^{\text{opt}} V_{jl} W_{0l}^{\text{opt}\top} = \left((I_p \otimes i_\tau)^\top V^{-1} (I_p \otimes i_\tau) \right)^{-1},$$

which is the smallest amongst our class of estimators.

We here give another interpretation of this procedure as an optimal minimum distance (omd) estimator described in [Rothenberg \(1973\)](#). This method arrives at the optimal combination of estimators through an explicit objective function. Let $\hat{\theta}_{\text{omd}}^\tau$ minimize the criterion function

$$Q_n(\theta) = \left[\begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_\tau \end{pmatrix} - \theta \otimes i_\tau \right]^\top V^{-1} \left[\begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_\tau \end{pmatrix} - \theta \otimes i_\tau \right], \quad (5.2)$$

where i_τ is a $\tau \times 1$ vector of ones, and V is the $\tau p \times \tau p$ asymptotic (as $n \rightarrow \infty$ holding τ constant) variance matrix of the vector $(\sqrt{n}(\hat{\theta}_1 - \theta_0)^\top, \dots, \sqrt{n}(\hat{\theta}_\tau - \theta_0)^\top)^\top$, i.e., $V = (V_{j,l})$. The first order condition

$$(I_p \otimes i_\tau)^\top V^{-1} \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_\tau \end{pmatrix} = (I_p \otimes i_\tau)^\top V^{-1} (\hat{\theta} \otimes i_\tau)$$

implies that the optimal minimum distance estimator $\hat{\theta}_{\text{omd}}^\tau$ is a linear combination of the $\hat{\theta}_j$ with

$$\hat{\theta}_{\text{omd}}^\tau = \sum_{j=1}^{\tau} W_{0j}^{\text{opt}} \hat{\theta}_j, \quad (5.3)$$

where the optimal weights are defined in (5.1).

Example 1 (*cont.*) Recall the optimal GMM estimator in this model (i.e., under homoskedasticity, etc.) is simply the two stage least squares estimator

$$\tilde{\theta} = (Y_2^\top P_X Y_2)^{-1} Y_2^\top P_X Y_1,$$

where $P_X = X(X^\top X)^{-1}X^\top$, $Y_1 = (y_{11}, \dots, y_{1n})^\top$, $Y_2 = (y_{21}, \dots, y_{2n})^\top$, $X = (X_1^\top, \dots, X_n^\top)$, $X_i = (X_{1i}, \dots, X_{ki})^\top$. Within our class of estimators \mathcal{E} , the optimal estimator is

$$\hat{\theta} = \sum_{j=1}^k W_{nj}^{\text{opt}} \hat{\theta}_j = (i_k^\top V^{-1} i_k)^{-1} i_k^\top V^{-1} \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_k \end{bmatrix},$$

where $\hat{\theta}_j = (Y_2^\top P_j Y_2)^{-1} Y_2^\top P_j Y_1$ for $j = 1, \dots, k$, where $P_j = X_j(X_j^\top X_j)^{-1}X_j^\top$ and V is the $k \times k$ covariance matrix with $V_{jl} = \text{asy. cov}(\hat{\theta}_j, \hat{\theta}_l)$. Suppose that the instruments are mutually orthogonal, then it is easy to see that $\hat{\theta}$ is identically equal to $\tilde{\theta}$.¹⁰ This gives yet another interpretation to 2SLS as being the optimal combination of exactly identified instrumental variables estimators.¹¹ Furthermore, $\hat{\theta}$ is computationally feasible even when $k > n$.¹²

5.1.1 Semiparametric Instrumental Variables

Suppose that we know only that

$$E[A_j(X_i)\rho(Z_i, \theta_0)] = 0, \quad j = 1, \dots, \tau, \quad (5.4)$$

where τ is fixed, and $A_j \in \mathbb{R}^p$. This is a standard unconditional moments estimation problem, and the optimal estimator (smallest variance) can be arrived at by several routes:

GMM with optimal combination of the moment conditions:

That is, we minimize the quadratic form

$$G_n^\tau(\theta)^\top \mathcal{W}_n G_n^\tau(\theta) \quad (5.5)$$

with respect to θ , where $G_n^\tau(\theta) = n^{-1} \sum_{i=1}^n A^\tau(X_i)\rho(Z_i, \theta)$ with $A^\tau = (A_1^\top, \dots, A_\tau^\top)^\top \in \mathbb{R}^{\tau p}$ (i.e., $G_n^\tau(\theta)$ is the $\tau p \times 1$ vector containing all the sample moments). The weighting matrix \mathcal{W}_n is such that $\mathcal{W}_n \rightarrow \mathcal{W}_{\text{opt}} \equiv \Psi_\tau^{-1}$ is the asymptotically optimal (opt) weighting matrix where $\Psi_\tau = E[G_n^\tau(\theta_0)G_n^\tau(\theta_0)^\top] = E[A^\tau(X)\sigma_0^2(X)A^\tau(X)^\top] \in \mathbb{R}^{\tau p \times \tau p}$.

Optimal instrumental variables:

¹⁰We are grateful to Tom Rothenberg for pointing this out to us.

¹¹Interpreting 2SLS in various ways has a long history in econometrics; see [Rothenberg \(1974\)](#) for an early example.

¹²This problem known as undersized sample problem arises in almost all large macroeconomic models with time series, because there are usually more pre-determined variables (lags) than time periods of observations (see [Klein, 1973](#) for an early account of alternative methods to solve this problem).

The optimal instrument in this case is simply a linear combination of the $A_j(X_i)$, $j = 1, \dots, \tau$. That is, we solve the equations

$$\Gamma^{\tau\top} \Psi_{\tau}^{-1} G_n^{\tau}(\widehat{\theta}) = 0, \quad (5.6)$$

where $\Gamma^{\tau} = \partial E[G_n^{\tau}(\theta_0)]/\partial\theta = E[A^{\tau}(X)D_0(X)^{\top}] \in \mathbb{R}^{\tau p \times p}$. These two approaches provide the oiv (optimal GMM) estimator $\widetilde{\theta}_{\text{oiv}}^{\tau}$ of θ_0 for the model (5.4). Specifically, we have $\sqrt{n}(\widetilde{\theta}_{\text{oiv}}^{\tau} - \theta_0) \implies N(0, \Sigma_{\text{oiv}}^{\tau})$ as $n \rightarrow \infty$, where the asymptotic variance is given by (see e.g., Hansen (1982) for differentiable ρ , Newey and McFadden (1994) for non-differentiable ρ):

$$\begin{aligned} \Sigma_{\text{oiv}}^{\tau} &= \left(E[A^{\tau}(X)D_0(X)^{\top}]^{\top} [E(A^{\tau}(X)\sigma_0^2(X)A^{\tau}(X)^{\top})]^{-1} E[A^{\tau}(X)D_0(X)^{\top}] \right)^{-1} \\ &= (\Gamma^{\tau\top} \Psi_{\tau}^{-1} \Gamma^{\tau})^{-1} \end{aligned} \quad (5.7)$$

and the optimal instrument for the model (5.4) is:

$$A_{\text{oiv}}^{\tau}(x) = \Gamma^{\tau\top} \Psi_{\tau}^{-1} A^{\tau}(x) = E[A^{\tau}(X)D_0(X)^{\top}]^{\top} [E(A^{\tau}(X)\sigma_0^2(X)A^{\tau}(X)^{\top})]^{-1} A^{\tau}(x).$$

In this case, we can show the equivalence between the optimal minimum distance estimator, as defined above, and the optimal instrumental variable estimator in the following proposition:

Proposition 1 *For each fixed τ , $\widehat{\theta}_{\text{omd}}^{\tau}$ is asymptotically efficient for (5.4) with $\Sigma_{\text{omd}}^{\tau} = \Sigma_{\text{oiv}}^{\tau}$. Moreover the optimal weighting is simply*

$$W_{0j}^{\text{oiv}} = - \left(\sum_{j=1}^{\tau} \alpha_j \Gamma_j^{\top} \right)^{-1} \alpha_j \Gamma_j^{\top} \text{ for } j = 1, \dots, \tau, \text{ with } (\alpha_1, \dots, \alpha_{\tau}) = \Gamma^{\tau\top} \Psi_{\tau}^{-1}.$$

5.2 Case 2: Increasing τ

Here we consider the more general case where τ increases with sample size. Notice that provided B5(a) is satisfied, the matrices $[(I_p \otimes i_{\tau})^{\top} V^{-1} (I_p \otimes i_{\tau})]^{-1}$ converge to a positive definite limit $\lim_{\tau \rightarrow \infty} \Sigma^{\tau}$. We next consider the important special case where a semiparametric efficiency standard is known against which we may compare our procedure.

5.2.1 Semiparametric Instrumental Variables

Let Σ_{oiv} be the asymptotic variance of the optimal instrumental variable (oiv) estimator, and let Σ_{omd} be the asymptotic variance as $n \rightarrow \infty$ and $\tau(n) \rightarrow \infty$ of the optimal minimum distance (omd) estimator defined above. The next theorem establishes the asymptotic equivalence between Σ_{omd} and Σ_{oiv} in the special case of an orthonormal basis.

ASSUMPTION C:

(C1) The matrix $D_0(X_i) = \left(\frac{\partial}{\partial \theta'} E[\rho(Z_i, \theta) | X_i]\right) |_{\theta=\theta_0}$ exists with probability one.

(C2) $E[\sigma_0^{-2}(X_i) D_0(X_i) D_0(X_i)^\top]$ is finite and positive definite.

(C3) $D_0(X_i) = \sum_{j=1}^{\infty} \beta_{j0} \phi_j(X_i) \sigma_0^2(X_i)$, where the sequence $\{\phi_j\}$ is a complete orthonormal basis satisfying:

$$E[\sigma_0^2(X_i) \phi_j(X_i) \phi_l(X_i)^\top] = \begin{cases} 0_p & \text{for } j \neq l, \\ I_p & \text{for } j = l. \end{cases}$$

Assumptions **C1** and **C2** are standard in the literature of efficient Instrumental Variable estimation. The first part of Assumption **C3** requires that the optimal matrix can be represented, in an L_2 , sense by a series expansion, and it has been used elsewhere, see e.g. [Carrasco and Florens \(2010\)](#) and examples therein. In the univariate case, the second part of this assumption can always be shown to hold in the homoskedastic case with $\sigma_0^2(X_i) = 1$ by letting $\{\psi_j\}$ be an orthonormal basis on the unit interval, and setting $\phi_j(x) = \psi_j(F(x))$, where F represents the CDF of X .¹³

Theorem 3 *Suppose that $E[\rho(Z_i, \theta_0) | X_i] = 0$ and that Assumptions **C1–C3** hold. Then,*

$$\Sigma_{\text{omd}} = \Sigma_{\text{oiv}} = \left(E[\sigma_0^{-2}(X_i) D_0(X_i) D_0(X_i)^\top]\right)^{-1}.$$

The optimal weights in this case are any sequence like

$$W_{nj}^0 = \left(\sum_{l=1}^{\tau(n)} V_{ll}^{-1}\right)^{-1} V_{jj}^{-1}, \quad (5.8)$$

where V_{jj} is the asymptotic variance matrix of $\sqrt{n}(\hat{\theta}_j - \theta_0)$. With such a sequence of weights, $\hat{\theta}$ has the same asymptotic variance as a comparable implementation of $\tilde{\theta}$. Note that in the scalar homoskedastic case, the optimal weights W_{nj}^0 , decrease at the same rate as β_{j0}^2 as $j \rightarrow \infty$, while the weights on the basis terms in the estimation of D_0 would decrease like β_{j0} as $j \rightarrow \infty$. This suggests that one needs to combine fewer estimators than instruments to achieve a specified variance.

6 Estimation of Optimal Weights

In this section, we consider estimation of the optimal weights and construction of a feasible asymptotically optimal estimator (in the sense of minimizing asymptotic variance within our class of procedures). In particular, we shall estimate the optimal weights defined in (5.1). Recall that V is the

¹³In this case, $E[\phi_j(X_i) \phi_l^\top(X_i)] = \int \phi_j(x) \phi_l^\top(x) F(dx) = \int_0^1 \phi_j(F^{-1}(u)) \phi_l^\top(F^{-1}(u)) du$.

$\tau p \times \tau p$ asymptotic (as $n \rightarrow \infty$ with τ fixed) covariance matrix of the vector of estimators $\widehat{\theta}_j$, $j \in \mathcal{J}_n$. We estimate the optimal weights as follows

$$\widehat{W}_{0j}^{\text{opt}} = \left(\sum_{l=1}^{\tau(n)} \widehat{B}_l \right)^{-1} \widehat{B}_j = [(I_p \otimes i_\tau)^\top \widehat{V}^{-1} (I_p \otimes i_\tau)]^{-1} [(I_p \otimes i_\tau)^\top \widehat{V}^{-1}]_j, \quad (6.1)$$

where $(\widehat{B}_1, \dots, \widehat{B}_\tau) = (I_p \otimes i_\tau)^\top \widehat{V}^{-1}$, and \widehat{V} has (j, l) sub-matrix calculated using formulae (4.4)–(4.5) for $j, l = 2, \dots, \tau$ based on a preliminary root- n consistent estimator of θ .

We next provide a consistency result for the estimator defined using the estimated weights (6.1). The strategy is to verify condition B4(a) of Theorem 2 for the estimated weights; we are implicitly assuming that B4(b) and B5 hold, so that the infeasible optimal weights are well defined. If the other conditions of Theorem 2 are satisfied, which are about the moment conditions, then the estimator based on (6.1), is consistent with Theorem 2. We shall restrict attention to the case where g_j are all differentiable. Define for each θ in a neighborhood of θ_0 ,

$$\Gamma_{nj}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial g_j}{\partial \theta}(Z_i, \theta); \quad \Omega_{njl}(\theta) = \frac{1}{n} \sum_{i=1}^n g_j(Z_i, \theta) g_l(Z_i, \theta)^\top,$$

$\Gamma_j(\theta) = E\Gamma_{nj}(\theta)$ and $\Omega_{jl}(\theta) = E\Omega_{njl}(\theta)$. Then $V_{jl} = \Gamma_j^{-1}(\theta_0) \Omega_{jl}(\theta_0) \Gamma_l^{-1}(\theta_0)^\top$. We shall assume the following high level conditions.

ASSUMPTION D:

(D1) The matrix V is finite and nonsingular for every τ , and $\lambda_{\min}(V) = o(\tau^{-\rho})$ for some $\rho \geq 0$.

(D2) For some $\rho_1 \geq 0$, $\min_{j \in \mathcal{J}_n} \lambda_{\min}(\Gamma_j) = o(\tau^{-\rho_1})$.

(D3) For some sequence $\delta_n \rightarrow 0$ and some $\eta > 0$:

$$\max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n \sqrt{n}} \|\Gamma_{nj}(\theta) - \Gamma_j(\theta)\| = o_p(n^{-\eta}); \quad \max_{j, l \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n \sqrt{n}} \|\Omega_{njl}(\theta) - \Omega_{jl}(\theta)\| = o_p(n^{-\eta}).$$

Assumption D1 is easy to verify in the orthonormal case, but is a reasonable assumption to maintain more generally. It allows the $\tau p \times \tau p$ matrix V to become asymptotically singular. In Example 1, when the errors are independent of the instruments, we have

$$V_{jl} \propto \frac{\text{cov}(x_j, x_l)}{\text{cov}(x_j, y_2) \text{cov}(x_l, y_2)},$$

and there are a variety of schemes for the covariance matrix of $(y_2, x_1, \dots, x_{\tau(n)})$ that would support assumption [D1](#). Similar comments apply to [D2](#) in the sense that it is easy to find a variety of examples consistent with this assumption.

Assumption [D3](#) can be verified under some primitive conditions, along the lines of the discussion around [Theorem 2](#). If $\tau(n) = c \log n$ for positive finite c , then the conditions are easy to satisfy. In this case, V is of relatively small dimension compared with the sample size available for estimation.

Theorem 4 *Suppose that Assumptions [D](#) hold and that $\tau(n) = c \log n$ for positive finite c . Then conditions [B4\(a\)](#) of [Theorem 2](#) hold.*

For further issues surrounding estimating the optimal weights for similar problems, we refer the reader to [Newey \(1990\)](#) and [Koenker and Machado \(1999\)](#).

7 Monte Carlo

We consider two data generating processes (DGPs). The first one is adapted from [Newey \(1990\)](#) who consider an endogenous dummy variable model with the following specification:

$$\begin{aligned} Y_i &= \beta_{10} + \beta_{20}s_i + \varepsilon_i; \\ \text{DGP1: } s_i &= 1(\alpha_{10} + \alpha_{20}X_i + \eta_i > 0), \\ X_i &\sim N(0, 1); \quad \alpha_{10} = \alpha_{20} = \beta_{10} = \beta_{20} = 1, \end{aligned}$$

where the errors ε_i and η_i are generated as

$$\begin{bmatrix} \varepsilon_i \\ \eta_i \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \varphi \\ \varphi & 1 \end{bmatrix} \right), \quad (7.1)$$

in which $\varphi \in \{0.2, 0.5, 0.8\}$ indicate weak, medium and strong endogeneity respectively. The optimal instrument for s is $\pi(x) = \Pr[s = 1|X = x]$, which makes $D(x) = (1, \pi(x))^\top$.

Figures [1–4](#) report results for two estimators of β_{20} . The first estimator (Estimator 1) corresponds to [Newey's \(1990\)](#) and two versions of the proposed estimators (Estimators 2 and 3) are also included. To obtain all estimators, we use 4 different basis: Basis 1 corresponds to the Hermite polynomials computed via the recursion $A_{j+1}(x) = 2xA_j(x) - 2jA_{j-1}(x)$, where $A_1(x) = 1$, $A_2 = 2x$, etc; Basis 2 corresponds to Legendre polynomials obtained via the recursion $(j+1)A_{j+1}(x) = (2j+1)xA_j(x) -$

$jA_{j-1}(x)$, where $A_1 = 1$, $A_2 = x$, etc; Basis 3 corresponds to Laguerre polynomials obtained via the recursion $A_{j+1}(x) = (j+1)^{-1}[(2j+1-x)A_j(x) - jL_{j-1}(x)]$, where $A_1(x) = 1$, $A_2(x) = 1-x$, etc; Basis 4 corresponds to the basis $A_j(x) = [x/(1+|x|)]^{j-1}$. For each basis, Estimator 1 becomes

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_{10} \\ \tilde{\beta}_{20} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n s_i \\ \sum_{i=1}^n \hat{\pi}(X_i) & \sum_{i=1}^n \hat{\pi}(X_i)s_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n \hat{\pi}(X_i)Y_i \end{pmatrix},$$

$$\hat{\pi}(x) = \sum_{j=1}^{\tau} \hat{\gamma}_j A_j(x).$$

for series-based estimated weights $\hat{\gamma}_j$. Using the same bases, our estimator becomes

$$\hat{\beta} = \sum_{j=2}^{\tau} W_{nj} \hat{\beta}_j, \text{ where} \quad (7.2)$$

$$\hat{\beta}_j = \begin{pmatrix} \hat{\beta}_{10;j} \\ \hat{\beta}_{20;j} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n s_i \\ \sum_{i=1}^n A_j(X_i) & \sum_{i=1}^n A_j(X_i)s_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n A_j(X_i)Y_i \end{pmatrix}. \quad (7.3)$$

Estimator 2 is calculated using fixed weights $W_{nj} = (j^{-3}/\sum_{j=2}^{\tau} j^{-3})I_2$, and Estimator 3 uses a feasible version of weights defined in (5.8). Results are presented for $\tau = 2, \dots, 6$ and two sample sizes: $n \in \{250, 1000\}$ and 5000 replications.

Figures 1–4 display the finite sample behaviour of the simulated $\{\tilde{\beta}^{(s)}\}_{s=1}^{5000}$ and two versions of $\{\hat{\beta}^{(s)}\}_{s=1}^{5000}$ as functions of τ in the form of box plots for each of the 4 bases. These figures contrast the estimators' different bias and variance behaviour as τ changes. Overall, the results are qualitatively similar between Estimators 1, 2 and 3 in terms of Monte Carlo bias and dispersion for each τ and sample size across bases. However, certain features are noteworthy. For example, one can observe the different behaviour between Newey's (1990) estimator (Estimator 1) and ours (Estimators 2 and 3) with respect to τ . Estimator 1 shows a tradeoff between bias (larger) and variance (smaller) as τ increases for both sample sizes and across bases, specially in the strong endogeneity case. On the other hand, only the variance seems to be affected by the choice of τ . Depending on the basis, the weights used and degree of endogeneity, τ equal to 2 or 3 would minimize the Monte Carlo mean square error for $n = 250$, and between 3 and 4 for $n = 1000$. Similarly, the proposed estimator seems to display smaller Monte Carlo dispersion when using Bases 2 and 4. Finally, although both versions of the proposed estimator perform comparably, its behavior seems to be more robust to the choice of τ when using fixed weights.

In DGP2 we consider a high-dimensional problem involving a two-equation system with the

following specification:

$$\begin{aligned}
 Y_i &= \beta_{10} + \beta_{20}s_i + \varepsilon_i; \\
 \text{DGP2: } s_i &= \alpha_{0;0} + \sum_{l=1}^k \alpha_{l;0}X_{li} + \eta_i, \\
 X_i &\sim N(0, I_k); \quad \alpha_{0;0} = \alpha_{1;0} = \dots = \alpha_{k;0} = \beta_{10} = \beta_{20} = 1,
 \end{aligned}$$

where $X_i = (X_{1i}, \dots, X_{ki})^\top$ and $(\varepsilon_i, \eta_i)^\top$ are generated as in (7.1). We set $k = 30$ and assess the performance of the ‘optimal’ estimator here in another set of 5000 replications, i.e. weights determined by (5.8) and $\tau = k$, with undersized samples of $n = 15$ and 25. Table 1 shows the Monte Carlo bias (Bias), standard deviations (Std. Dev) and root mean squared error (RMSE). The new estimator is compared against generic 2SLS. Notice that for these sample sizes generic 2SLS with as many instruments as the sample size is equivalent to Ordinary Least Squares (OLS).¹⁴ The proposed estimator shows small biases and decreasing (with respect to sample size) variances for each endogeneity parameter value, while the OLS estimator displays considerably larger bias as expected. However, for a sample size of 50 observations, the variance and RMSE of the proposed estimator are comparable to that of 2SLS.

For illustration purposes, Figure 5 displays the estimated $\widehat{\beta}_{20; j}$ in (7.3) for $j = 1, \dots, 9$ using Basis 3 in DGP 1 for a sample size of $n = 1000$ with $\rho = 0.8$ against their standard errors (gray points). The dotted line represents the true value $\beta_{20} = 1$. One can observe the large tradeoff between bias and variance depending on the instrument being used. The same plot also displays the estimated $\widehat{\beta}$ in (7.2) for $\tau = 2, \dots, 9$ with weights given by a feasible version of (5.8), against their standard errors (black points). In this case both bias and variance of the proposed estimator decays when combining the first 3 estimates, but they increase when adding further elements in the average. Unlike $\widehat{\beta}_j$ the tradeoffs between bias and variance is less dramatic for the combined estimator.

8 Conclusions, Extensions and Practical Considerations

This paper provides a new way of calculating efficient semiparametric instrumental variable estimators by means of constructing linear combinations of an infinite dimensional set of \sqrt{n} -consistent but inefficient estimators, where the cardinality of the linear combination increases with sample size. Our

¹⁴Although alternative methods that rely on choosing a subset of instruments are readily available, see e.g. Donald and Newey (2001) and Kuersteiner and Okui (2010).

approach has an advantage over the traditional approach to semiparametric instrumental variables in that one has a ‘distribution’ of estimators of the same quantity and one can view the range of values that these estimators take. If that range is not great, then it would appear that achieving efficiency is not going to be worth very much. If the range is considerable, then the efficient estimator may be very much better than any given estimator but at the same time performance might be very sensitive to how it is constructed. This information contained in the spread of the different estimators is similar to but not necessarily the same as the information contained in the standard error of an efficient estimator.¹⁵ Also, the optimal weighting just requires the estimation of HAC matrices, at least in the orthonormal basis case, about which much has been written in econometrics.

It is quite straightforward to extend our work to produce results for the range¹⁶

$$\mathfrak{R}_n = \max_{j \in \mathcal{J}_n} \widehat{\theta}_j - \min_{j \in \mathcal{J}_n} \widehat{\theta}_j$$

using the theory of extreme values for Gaussian processes (as in [Bickel and Breiman, 1983](#), for example). This statistic can be used as another way of measuring whether the observed range is consistent with the underlying model assumptions, i.e. as a model specification test.

Recently, [Lobato and Dominguez \(2004\)](#) has pointed out that global identification of models defined by conditional moment restrictions can be lost when using a set of unconditional moments (even when constructed with the optimal instruments). The proposed methodology can potentially be used to recover both global identification (by means of using a set of inefficient but valid instruments) as well as efficiency (via an optimal combination as discussed above).

Practical Choice of Weights, \mathcal{J}_n and τ

We have shown how to compute an optimal estimator for a given choice of \mathcal{J}_n . The theoretical analysis of methods for determining \mathcal{J}_n is quite complex and would justify a separate paper, since it involves a higher order theory. In theory, the larger is \mathcal{J}_n the better in terms of variance, but in practice there is a tradeoff. Let $\widehat{\theta}(\mathcal{J}_n)$ be the feasible optimal estimator as computed in the previous section, and let $\widehat{\Sigma}_{\text{opt}}(\mathcal{J}_n)$ be a consistent estimator of its asymptotic variance. Along the lines of [Politis and Romano \(1992\)](#), one could choose \mathcal{J}_n to be the place where the standard errors (a scalar

¹⁵Actually, if the estimators themselves are mutually independent with the same limiting distribution, then the 95% confidence interval of a single estimator is approximately the same as the inter hemi-decile range, that is the interval $[\widehat{\theta}_{0.025 \cdot \tau}, \widehat{\theta}_{0.975 \cdot \tau}]$ of the ordered estimators. In fact, it is not possible that the estimators come from the same asymptotic distribution (since the variances must diverge along some trajectory), and so the two intervals do not coincide. Nevertheless, the connection exists.

¹⁶In the multivariate case, we take the coordinate-wise ranges.

function of the covariance matrix) are relatively stable and do not vary wildly as \mathcal{J}_n varies close by. In practice, we have found it useful to plot the input estimators against their marginal standard errors as an informal device to select reasonable estimators, see e.g. Figure 5 in the previous section.

We now conclude by describing how one would select the set \mathcal{J}_n in a given application with finite sample n . There are a number of informal methods a practitioner can use. For example, for given τ one could compute the t-statistic (in the scalar case, otherwise a chi-squared statistic) for each estimator and retain only those estimators with the largest τ such values. Alternatively, one could choose to retain only those estimators with t-statistic, say, that exceeds some predetermined level like one. Alternatively, if G_{nj} is continuously differentiable with respect to θ , one could also use results in [Rilstone, Srivastava, and Ullah \(1996\)](#); [Rilstone and Ullah \(2005\)](#) to estimate \mathcal{J}_n and the weights by means of minimizing an estimate of the proposed estimator's mean square error, see e.g. [Schafgans and Zinde-Walsh \(2010\)](#). The theoretical justification of the latter is left for future research.

References

- AMEMIYA, T. (1974): "The nonlinear two-stage least-squares estimator," *Journal of Econometrics*, 2(2), 105–110.
- BICKEL, P. J., AND L. BREIMAN (1983): "Sums of Functions of Nearest Neighbor Distances, Moment Bounds, Limit Theorems and a Goodness of Fit Test," *The Annals of Probability*, 11(1), 185–214.
- BIERENS, H. J. (1987): "Kernel Estimators of Regression Functions," in *Advances in Econometrics – Fifth World Congress of the Econometric Society*, ed. by T. F. Bewley, vol. I of *Econometric Society Monographs*, chap. 3, pp. 99–144. Cambridge University Press, 1 edn.
- BRAVO, F., J. C. ESCANCIANO, AND T. OTSU (2011): "A Simple Test for Identification in GMM under Conditional Moment Restrictions," Discussion Paper 1790, Cowles Foundation.
- BREIMAN, L. (1996): "Bagging Predictors," *Machine Learning*, pp. 123–140.
- (1999): "Using adaptive bagging to debias regressions," Technical Report 547, Department of Statistics, University of California Berkeley.
- CAMPBELL, J. Y., A. LO, AND A. C. MACKINLAY (1997): *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey, 1 edn.

- CARRASCO, M., AND J.-P. FLORENS (2010): “On the Asymptotic Efficiency of GMM,” Unpublished Manuscript.
- CHAMBERLAIN, G. (1987): “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions,” *Journal of Econometrics*, 34, 305–334.
- CHEN, X., AND D. POUZO (2009): “Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals,” *Journal of Econometrics*, 152(1), 46–60.
- DONALD, S. G., G. W. IMBENS, AND W. K. NEWEY (2003): “Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions,” *Journal of Econometrics*, 117(1), 55–93.
- DONALD, S. G., AND W. K. NEWEY (2001): “Choosing the Number of Instruments,” *Econometrica*, 69(5), 1161–91.
- GRANGER, C. W. J., AND Y. JEON (2004): “Thick modeling,” *Economic Modelling*, 21(2), 323–343.
- GRAY, H. L., AND W. R. SCHUCANY (1972): *The Generalized Jackknife Statistic*. Marcel Dekker, New York, 1 edn.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Methods of Moments Estimators,” *Econometrica*, 50, 1029–1054.
- (1985): “A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators,” *Journal of Econometrics*, 30(1-2), 203–238.
- HUBER, P. J. (1967): “The behavior of maximum likelihood estimates under nonstandard conditions,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, ed. by L. M. L. Cam, and J. Neyman, pp. 221–233, Statistical Laboratory of the University of California, Berkeley. University of California Press.
- INOUE, A., AND B. ROSSI (2011): “Testing for weak identification in possibly nonlinear models,” *Journal of Econometrics*, 161(2), 246–261.
- KLEIN, L. (1973): “The Treatment of Undersize Samples in Econometrics,” in *Econometrics Studies of Macro and Monetary Relations*, ed. by A. A. Powell, and R. A. Williams, pp. 3–26. American Elsevier Publishing Company, Inc., North-Holland, Amsterdam, 1 edn.

- KOENKER, R., AND J. A. F. MACHADO (1999): “GMM inference when the number of moment conditions is large,” *Journal of Econometrics*, 93(2), 327–344.
- KOENKER, R. W., AND J. BASSETT, GILBERT (1978): “Regression Quantiles,” *Econometrica*, 46(1), 33–50.
- KOTLYAROVA, Y., AND V. ZINDE-WALSH (2006): “Non- and semi-parametric estimation in models with unknown smoothness,” *Economics Letters*, 93(3), 379–386.
- (2007): “Robust kernel estimator for densities of unknown smoothness,” *Journal of Non-parametric Statistics*, 19(2), 89–101.
- KUERSTEINER, G., AND R. OKUI (2010): “Constructing Optimal Instruments by First-Stage Prediction Averaging,” *Econometrica*, 78(2), 697–718.
- LEE, Y., AND Y. ZHOU (2011): “Averaged Instrumental Variables Estimator,” Unpublished Manuscript.
- LOBATO, I. N., AND M. A. DOMINGUEZ (2004): “Consistent Estimation of Models Defined by Conditional Moment Restrictions,” *Econometrica*, 72(5), 1601–1615.
- MALINVAUD, E. (1966): *Statistical methods of econometrics*, Studies in Mathematical and Managerial Economics. Rand McNally, 1 edn.
- McFADDEN, D. (1989): “A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration,” *Econometrica*, 57(5), 995–1026.
- NEWKEY, W. K. (1990): “Efficient Instrumental Variables Estimation of Nonlinear Models,” *Econometrica*, 58(4), 809–37.
- NEWKEY, W. K. (1993): “Efficient estimation of models with conditional moment restrictions,” in *Handbook of Statistics*, vol. 11, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, pp. 419–454. Elsevier Publishing Company, Inc., North-Holland, Amsterdam, 1 edn.
- NEWKEY, W. K., AND D. McFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by D. McFadden, and R. F. Engle, vol. IV, pp. 2111–2245. Elsevier, North-Holland, Amsterdam.

- NEWHEY, W. K., AND R. J. SMITH (2004): “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72(1), 219–255.
- PAKES, A., AND D. POLLARD (1989): “Simulation and the Asymptotics of Optimization Estimators,” *Econometrica*, 57(5), 1027–57.
- PHILLIPS, P. C. B. (1983): “Exact small sample theory in the simultaneous equations model,” in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. 1 of *Handbook of Econometrics*, chap. 8, pp. 449–516. Elsevier, 1 edn.
- POLITIS, D. N., AND J. P. ROMANO (1992): “A General Resampling Scheme for Triangular Arrays of α -mixing Random Variables with Application to the Problem of Spectral Density Estimation,” *Annals of Statistics*, 20(4), 1985–2007.
- PORTNOY, S. (1984): “Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency,” *The Annals of Statistics*, 12(4), 1298–1309.
- RILSTONE, P., V. K. SRIVASTAVA, AND A. ULLAH (1996): “The second-order bias and mean squared error of nonlinear estimators,” *Journal of Econometrics*, 75(2), 369–395.
- RILSTONE, P., AND A. ULLAH (2005): “Corrigendum to ‘The second-order bias and mean squared error of nonlinear estimators’; [Journal of Econometrics 75(2) (1996) 369-395],” *Journal of Econometrics*, 124(1), 203–204.
- ROTHENBERG, T. J. (1973): *Efficient Estimation with a priori Information*. Yale University Press, New Haven, USA, 1 edn.
- (1974): “A Note on Two-Stage Least Squares,” Unpublished manuscript.
- SAWA, T. (1973): “Almost Unbiased Estimator in Simultaneous Equations Systems,” *International Economic Review*, 14(1), 97–106.
- SCHAFGANS, M. M. A., AND V. ZINDE-WALSH (2010): “Smoothness adaptive average derivative estimation,” *Econometrics Journal*, 13(1), 40–62.
- STOCK, J. H., AND M. W. WATSON (1999): “Forecasting inflation,” *Journal of Monetary Economics*, 44(2), 293–335.

STOCK, J. H., AND J. WRIGHT (2000): “GMM with Weak Identification,” *Econometrica*, 68(5), 1055–1096.

WATSON, M. W. (2003): “Macroeconomic Forecasting Using Many Predictors,” in *Advances in Economics and Econometrics Theory and Applications – Eight World Congress of the Econometric Society*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, vol. III of *Econometric Society Monographs*, chap. 3, pp. 87–115. Cambridge University Press, 1 edn.

Appendix A: Mathematical Proofs

Proof of Lemma 1. We show that

$$\Pr \left[\max_{j \in \mathcal{J}_n} \left| \sum_{i=1}^n U_{ji} \right| \geq \lambda_n \right] \rightarrow 0$$

for any $\lambda_n = \delta_n \sqrt{n}$. For an array $\chi_{nj} \rightarrow \infty$ as $n \rightarrow \infty$ for each j , write

$$U_{ji} = U_{ji}1(|U_{ji}| \leq \chi_{nj}) + U_{ji}1(|U_{ji}| > \chi_{nj}) = \tilde{U}_{ji} + \tilde{\tilde{U}}_{ji}.$$

We shall assume for simplicity that U_{ji} is symmetric about zero so that $E(\tilde{U}_{ji}) = 0$. Therefore, \tilde{U}_{ji} are i.i.d. for each j with mean zero and are bounded from above by χ_{nj} . By the Bonferroni and Bernstein inequalities

$$\begin{aligned} \Pr \left[\max_{j \in \mathcal{J}_n} \left| \sum_{i=1}^n \tilde{U}_{ji} \right| \geq \lambda_n \right] &\leq \sum_{j \in \mathcal{J}_n} \Pr \left[\left| \sum_{i=1}^n \tilde{U}_{ji} \right| \geq \lambda_n \right] \\ &\leq \sum_{j \in \mathcal{J}_n} \exp \left(\frac{-\lambda_n^2}{s_{nj}^2 + 2\lambda_n \chi_{nj}} \right). \end{aligned} \tag{A-1}$$

We shall choose λ_n and χ_{nj} below to make this term vanish.

By the Bonferroni and Markov inequalities

$$\begin{aligned}
\Pr \left[\max_{j \in \mathcal{J}_n} \left| \sum_{i=1}^n \tilde{U}_{ji} \right| \geq \lambda_n \right] &\leq \sum_{j \in \mathcal{J}_n} \Pr \left[\left| \sum_{i=1}^n \tilde{U}_{ji} \right| \geq \lambda_n \right] \\
&\leq \sum_{j \in \mathcal{J}_n} \frac{E \left(\left| \sum_{i=1}^n \tilde{U}_{ji} \right|^\kappa \right)}{\lambda_n^\kappa} \\
&\leq \sum_{j \in \mathcal{J}_n} \frac{n^\kappa E(|U_{ji}|^\kappa) \Pr[|U_{ji}| > \chi_{nj}]}{\lambda_n^\kappa} \\
&\leq \sum_{j \in \mathcal{J}_n} \frac{n^\kappa [E(|U_{ji}|^\kappa)]^2}{\lambda_n^\kappa \chi_{nj}^\kappa} = o(1)
\end{aligned}$$

provided $\sum_{j \in \mathcal{J}_n} n^\kappa \chi_{nj}^{-\kappa} \lambda_n^{-\kappa} c_j^2 \rightarrow 0$.

Letting $\lambda_n = \delta_n \sqrt{n}$ and $\chi_{nj} = \sigma_j^2 \sqrt{n}$ we need to show that:

$$\sum_{j \in \mathcal{J}_n} \exp \left(\frac{-\delta_n}{\sigma_j^2} \right) \rightarrow 0 \text{ and } \frac{1}{\delta_n^\kappa} \sum_{j \in \mathcal{J}_n} \frac{c_j^2}{\sigma_j^{2\kappa}} \rightarrow 0.$$

For the first condition it suffices that

$$\frac{\delta_n}{\max_{j \in \mathcal{J}_n} \sigma_j^2 \log \tau(n)} \rightarrow \infty.$$

For the second condition it certainly suffices if

$$\frac{\delta_n}{\left(\sum_{j \in \mathcal{J}_n} c_j^2 \sigma_j^{-2\kappa} \right)^{1/\kappa}} \rightarrow \infty.$$

Proof of Theorem 1 (i). From [A2](#), if $\max_{j \in \mathcal{J}_n} \|\hat{\theta}_j - \theta_0\| > \delta$, then $\|G_j(\hat{\theta}_j)\| \geq \epsilon_n(\delta)$ for some j . Consequently

$$\Pr \left(\max_{j \in \mathcal{J}_n} \|\hat{\theta}_j - \theta_0\| > \delta \right) \leq \Pr \left(\max_{j \in \mathcal{J}_n} \|G_j(\hat{\theta}_j)\| \geq \epsilon_n(\delta) \right), \tag{A-2}$$

and it is sufficient to prove that for the given $\epsilon_n(\delta) > 0$, the latter probability goes to zero. But

$$\begin{aligned}
\max_{j \in \mathcal{J}_n} \|G_j(\widehat{\theta}_j)\| &\leq \max_{j \in \mathcal{J}_n} \|G_j(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)\| + \max_{j \in \mathcal{J}_n} \|G_{nj}(\widehat{\theta}_j)\| \text{ by the Triangle Inequality,} \\
&\leq \max_{j \in \mathcal{J}_n} \sup_{\theta \in \Theta} \|G_j(\theta) - G_{nj}(\theta)\| + \max_{j \in \mathcal{J}_n} \|G_{nj}(\widehat{\theta}_j)\| \text{ by set inclusion,} \\
&= o_p(\alpha_{2n}) + \max_{j \in \mathcal{J}_n} \|G_{nj}(\widehat{\theta}_j)\| \text{ by A4,} \\
&\leq o_p(\alpha_{2n}) + \max_{j \in \mathcal{J}_n} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\theta \in \Theta} \|G_{nj}(\theta)\| \right) + \max_{j \in \mathcal{J}_n} \inf_{\theta \in \Theta} \|G_{nj}(\theta)\|, \\
&\leq o_p(\alpha_{2n}) + \max_{j \in \mathcal{J}_n} \left(\|G_{nj}(\widehat{\theta}_j)\| - \inf_{\theta \in \Theta} \|G_{nj}(\theta)\| \right) + \max_{j \in \mathcal{J}_n} \|G_{nj}(\theta_0)\|, \\
&= o_p(\alpha_{2n}) + o_p(\alpha_{1n}) = o_p(\epsilon_n(\delta))
\end{aligned}$$

by A3, A4 and the definition of θ_0 . We conclude that $\max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| = o_p(1)$. Finally,

$$\|\widehat{\theta} - \theta_0\| \leq \sum_{j \in \mathcal{J}_n} \|W_{nj}\| \times \max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| = o_p(1)$$

by A1. ■

Proof of Theorem 1 (ii). Consistency Theorem 1 (i) implies that for every $\epsilon > 0$ there exists a sequence $\{\delta_n\}$ with $\delta_n \rightarrow 0$, and an N such that for all $n \geq N$, $\Pr\{\max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| > \delta_n\} \leq \epsilon$. Using the same proof as that of Theorem 1 (i), we have under our stronger assumption A*4 that with probability approaching 1 (wpa1)

$$\begin{aligned}
\max_{j \in \mathcal{J}_n} \|G_j(\widehat{\theta}_j)\| &\leq \max_{j \in \mathcal{J}_n} \|G_j(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)\| + \max_{j \in \mathcal{J}_n} \|G_{nj}(\widehat{\theta}_j)\| \\
&\leq \max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq \delta_n} \|G_j(\theta) - G_{nj}(\theta)\| + \max_{j \in \mathcal{J}_n} \|G_{nj}(\widehat{\theta}_j)\| \\
&= o_p(\epsilon_n n^{-1/4}) + \max_{j \in \mathcal{J}_n} \|G_{nj}(\widehat{\theta}_j)\| \text{ by A*4,} \\
&= o_p(\epsilon_n n^{-1/4}) \text{ by A*3, A*4 and the definition of } \theta_0.
\end{aligned}$$

Therefore, by A*2

$$\max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| \leq \frac{1}{\min_{j \in \mathcal{J}_n} \gamma_j} \max_{j \in \mathcal{J}_n} \|G_j(\widehat{\theta}_j)\| = o(n^{-1/4}).$$

Hence

$$\max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| = o_p(n^{-1/4}),$$

which implies that

$$\|\widehat{\theta} - \theta_0\| \leq \sum_{j \in \mathcal{J}_n} \|W_{nj}\| \times \max_{j \in \mathcal{J}_n} \|\widehat{\theta}_j - \theta_0\| = o_p(n^{-1/4})$$

as required, since $\sum_{j \in \mathcal{J}_n} \|W_{nj}\|$ is uniformly bounded by [A1](#). ■

Proof of Theorem 2. Let

$$L_{nj}(\theta) = G_{nj}(\theta_0) + \Gamma_j(\theta - \theta_0)$$

for each $j = 1, 2, \dots$. Then define θ_j^* as the minimizer of $\|L_{nj}(\theta)\|$ over $\theta \in R^p$ (Note that θ_j^* minimizes over R^p , and not over Θ . We ignore this difference below because θ_j^* will eventually be in Θ w.p.1.). The solution satisfies

$$\sqrt{n}(\theta_j^* - \theta_0) = -\Gamma_j^{-1} \sqrt{n} G_{nj}(\theta_0) \tag{A-3}$$

for each j . Therefore,

$$\begin{aligned} \sqrt{n} \sum_{j \in \mathcal{J}_n} W_{nj}(\theta_j^* - \theta_0) &= \sqrt{n} \sum_{j \in \mathcal{J}_n} W_{nj}^0(\theta_j^* - \theta_0) + \sqrt{n} \sum_{j \in \mathcal{J}_n} (W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0) \\ &= \sum_{i=1}^n T_{in} + R_n, \end{aligned}$$

where $R_n = \sqrt{n} \sum_{j \in \mathcal{J}_n} (W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0)$ and $T_{in} = \frac{-1}{\sqrt{n}} \sum_{j \in \mathcal{J}_n} W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0)$.

The result follows after we establish:

- (i) $\sum_{i=1}^n c^\top T_{in} \implies N(0, c^\top \Sigma c)$ for any $c \in R^p$ with $\|c\| = 1$;
- (ii) The remainder term $R_n = o_p(1)$;
- (iii) $\sqrt{n} \sum_{j \in \mathcal{J}_n} W_{nj}(\theta_j^* - \widehat{\theta}_j) = o_p(1)$.

For (i), the triangular array of random variables $c^\top T_{in}$ is mean zero and independent across i for each n . By [B5\(a\)](#) we have:

$$\begin{aligned} \sum_{i=1}^n E[c^\top T_{in}]^2 &= E \left[\left(\sum_{j \in \mathcal{J}_n} c^\top W_{nj}^0 \Gamma_j^{-1} g_j(Z_i, \theta_0) \right)^2 \right] \\ &= \sum_{j \in \mathcal{J}_n} \sum_{l \in \mathcal{J}_n} c^\top W_{nj}^0 \Gamma_j^{-1} E [g_j(Z_i, \theta_0) g_l(Z_i, \theta_0)^\top] \Gamma_l^{-1 \top} W_{nl}^{0 \top} c \\ &\rightarrow c^\top \Sigma c . \end{aligned}$$

Similarly, by **B5(b)** we have for some $\kappa > 0$,

$$\sum_{i=1}^n E|c^\top T_{in}|^{2+\kappa} \rightarrow 0.$$

Hence we obtain (i) by applying the Liapunov's triangular array central limit theorem.

For (ii), notice that Assumption **B3(a)** and **(A-3)** imply that $\max_{j \in \mathcal{J}_n} \|\Gamma_j \sqrt{n}(\theta_j^* - \theta_0)\| = O_p(1)$. This together with Assumption **B4(a)** imply (ii) because

$$\begin{aligned} \left\| \sqrt{n} \sum_{j \in \mathcal{J}_n} (W_{nj} - W_{nj}^0)(\theta_j^* - \theta_0) \right\| &\leq \sqrt{n} \max_{j \in \mathcal{J}_n} \|\Gamma_j(\theta_j^* - \theta_0)\| \sum_{j \in \mathcal{J}_n} \|(W_{nj} - W_{nj}^0)\Gamma_j^{-1}\| \\ &= O_p(1) \times o_p(1). \end{aligned}$$

For (iii), by the $n^{1/4}$ -consistency result, there exists a positive sequence $\eta_n \rightarrow 0$ such that $\Pr[n^{1/4} \|\widehat{\theta} - \theta_0\| > \eta_n] \rightarrow 0$. For each j we have

$$\begin{aligned} G_{nj}(\theta) &= G_{nj}(\theta_0) + G_j(\theta) + G_{nj}(\theta) - G_j(\theta) - G_{nj}(\theta_0) \\ &= L_{nj}(\theta) + O(\|\theta - \theta_0\|^2) + [G_{nj}(\theta) - G_j(\theta)] - G_{nj}(\theta_0) \text{ by } \mathbf{B2}. \end{aligned}$$

Therefore, for the above η_n and constants a and C we have

$$\begin{aligned} \max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq a\eta_n/n^{1/4}} \sqrt{n} \|G_{nj}(\theta) - L_{nj}(\theta)\| \\ \leq C \times \eta_n^2 a^2 + \max_{j \in \mathcal{J}_n} \sup_{\|\theta - \theta_0\| \leq a\eta_n/n^{1/4}} \sqrt{n} \|[G_{nj}(\theta) - G_j(\theta)] - G_{nj}(\theta_0)\| \\ = O_p(\eta_n^2) + o_p(1) = o_p(1) \text{ by } \mathbf{B3(b)}. \end{aligned}$$

Therefore,

$$\max_{j \in \mathcal{J}_n} \|\sqrt{n}[L_{nj}(\theta_j^*) - G_{nj}(\theta_j^*)]\| = o_p(1), \text{ and } \max_{j \in \mathcal{J}_n} \|\sqrt{n}[L_{nj}(\widehat{\theta}_j) - G_{nj}(\widehat{\theta}_j)]\| = o_p(1)$$

because θ_j^* is \sqrt{n} -consistent and $\widehat{\theta}_j$ is $o(n^{-1/4})$ -consistent. It now follows from the definition of θ_j^* and Assumption **B1** and the triangular inequality that

$$\max_{j \in \mathcal{J}_n} \left| \sqrt{n} \|L_{nj}(\theta_j^*)\| - \sqrt{n} \|L_{nj}(\widehat{\theta}_j)\| \right| = o_p(1). \quad (\text{A-4})$$

This implies that $\max_{j \in \mathcal{J}_n} \|\Gamma_j \sqrt{n}(\theta_j^* - \hat{\theta}_j)\| = o_p(1)$, because of the properties of least squares residuals. Then we have

$$\begin{aligned} \sqrt{n} \sum_{j \in \mathcal{J}_n} W_{nj}(\theta_j^* - \hat{\theta}_j) &\leq \sum_{j \in \mathcal{J}_n} \|W_{nj} \Gamma_j^{-1}\| \times \max_{j \in \mathcal{J}_n} \|\Gamma_j \sqrt{n}(\theta_j^* - \hat{\theta}_j)\| \\ &\leq O_p(1) \times o_p(1) = o_p(1), \end{aligned}$$

where the last inequality is due to Assumption B4(a) and (b) since

$$\begin{aligned} \sum_{j \in \mathcal{J}_n} \|W_{nj} \Gamma_j^{-1}\| &\leq \sum_{j \in \mathcal{J}_n} \|W_{nj}^0 \Gamma_j^{-1}\| + \sum_{j \in \mathcal{J}_n} \|(W_{nj} - W_{nj}^0) \Gamma_j^{-1}\| \\ &= O(1) + o_p(1) = O_p(1), \end{aligned}$$

the result (iii) follows. ■

Proof of Proposition 1. On the one-hand, by the results in Hansen (1982), the optimal GMM (oiv) estimator is asymptotically efficient among all regular \sqrt{n} -asymptotic normal estimators for the moment restrictions (5.4), hence $\Sigma_{\text{oiv}}^\tau \leq \Sigma_{\text{omd}}^\tau$ in the positive semi-definite matrix sense. On the other hand, we notice that the oiv (optimal GMM) estimator has the expansion

$$\sqrt{n}(\tilde{\theta}_{\text{oiv}}^\tau - \theta_0) = -(\Gamma^{\tau\top} \Psi_\tau^{-1} \Gamma^\tau)^{-1} \Gamma^{\tau\top} \Psi_\tau^{-1} \sqrt{n} G_n^\tau(\theta_0) + o_p(1),$$

which can be rewritten as

$$\sqrt{n}(\tilde{\theta}_{\text{oiv}}^\tau - \theta_0) = - \left(\sum_{j=1}^{\tau} \alpha_j \Gamma_j^\top \right)^{-1} \sum_{j=1}^{\tau} \alpha_j \sqrt{n} G_{nj}(\theta_0) + o_p(1), \quad (\text{A-5})$$

where $\Gamma^{\tau\top} \Psi_\tau^{-1} = (\alpha_1, \dots, \alpha_\tau)$ with $\alpha_j \in R^{p \times p}$, and $\Gamma^\tau = (\Gamma_1^\top, \dots, \Gamma_\tau^\top)^\top$ with $\Gamma_j = E[A_j(X) D_0(X)^\top]$, and $G_{nj}(\theta) = \frac{1}{n} \sum_{i=1}^n A_j(X_i) \rho(Z_i, \theta)$ for $j = 1, \dots, \tau$. That is, the optimal GMM (oiv) estimator $\tilde{\theta}_{\text{oiv}}^\tau$ belongs to the class of linear combinations of the $\hat{\theta}_j$, $j = 1, \dots, \tau$ with

$$\tilde{\theta}_{\text{oiv}}^\tau = \sum_{j=1}^{\tau} W_{0j}^{\text{oiv}} \hat{\theta}_j + o_p(n^{-1/2}),$$

and

$$W_{0j}^{\text{oiv}} = - \left(\sum_{j=1}^{\tau} \alpha_j \Gamma_j^\top \right)^{-1} \alpha_j \Gamma_j^\top \text{ for } j = 1, \dots, \tau.$$

However, by the results in Rothenberg (1973), $\hat{\theta}_{\text{omd}}^\tau = \sum_{j=1}^{\tau} W_{0j}^{\text{opt}} \hat{\theta}_j$ is asymptotically efficient among the regular class of estimators of the form $\sum_{j=1}^{\tau} W_{0j} \hat{\theta}_j$ with $\sum_{j=1}^{\tau} W_{0j} = I_p$, hence $\Sigma_{\text{omd}}^\tau \leq \Sigma_{\text{oiv}}^\tau$ in the positive semi-definite matrix sense. Therefore $\Sigma_{\text{omd}}^\tau = \Sigma_{\text{oiv}}^\tau$ in (5.7). ■

Proof of Theorem 3. Assumption **C3** implies that $\beta_{j0} = E[D_0(X_i)\phi_j(X_i)^\top]$. We have:

$$\begin{aligned}\Sigma_{\text{oiv}} &= (E[\sigma_0^{-2}(X_i)D_0(X_i)D_0(X_i)^\top])^{-1} = \left(E \left[\sum_{j=1}^{\infty} \beta_{j0}\phi_j(X_i)\sigma_0^2(X_i)\sigma_0^{-2}(X_i)D_0(X_i)^\top \right] \right)^{-1} \\ &= \left(\sum_{j=1}^{\infty} \beta_{j0} E[\phi_j(X_i)D_0(X_i)^\top] \right)^{-1} = \left(\sum_{j=1}^{\infty} \beta_{j0}\beta_{j0}^\top \right)^{-1} \\ &= \left(\sum_{j=1}^{\infty} E[D_0(X_i)^\top\phi_j(X_i)^\top] E[\phi_j(X_i)D_0(X_i)] \right)^{-1}.\end{aligned}$$

Assumptions **C2** and **C3** imply that $0 < \sum_{j=1}^{\infty} \beta_{j0}\beta_{j0}^\top < \infty$.

By Assumptions **C1–C3**, we have $V_{jj} = \{\Gamma_j^\top \Gamma_j\}^{-1} = \{E[\phi_j(X_i)D_0(X_i)]^\top E[\phi_j(X_i)D_0(X_i)]\}^{-1}$ and $V_{jl} = 0$ for all $j \neq l$. Therefore

$$\begin{aligned}\Sigma_{\text{omd}} &= \lim_{\tau \rightarrow \infty} [(I_p \otimes i_\tau)^\top V^{-1}(I_p \otimes i_\tau)]^{-1} \\ &= \lim_{\tau \rightarrow \infty} \left(\sum_{j=1}^{\tau} V_{jj}^{-1} \right)^{-1} \\ &= \lim_{\tau \rightarrow \infty} \left(\sum_{j=1}^{\tau} \{E[\phi_j(X_i)D_0(X_i)]^\top E[\phi_j(X_i)D_0(X_i)]\} \right)^{-1} \\ &= \left(\sum_{j=1}^{\infty} \beta_{j0}\beta_{j0}^\top \right)^{-1}.\end{aligned}$$

■

Proof of Theorem 4. We have

$$\sum_{j \in \mathcal{J}_n} \|(\widehat{W}_{0j}^{\text{opt}} - W_{0j}^{\text{opt}})\Gamma_j^{-1}\| \leq \tau(n)^{1+\rho_1} \max_{j \in \mathcal{J}_n} \|(\widehat{W}_{0j}^{\text{opt}} - W_{0j}^{\text{opt}})\|,$$

where

$$\begin{aligned}\widehat{W}_{0j}^{\text{opt}} - W_{0j}^{\text{opt}} &= [(I_p \otimes i_\tau)^\top \widehat{V}^{-1}(I_p \otimes i_\tau)]^{-1} \widehat{B}_j - [(I_p \otimes i_\tau)^\top V^{-1}(I_p \otimes i_\tau)]^{-1} B_j \\ &= [(I_p \otimes i_\tau)^\top V^{-1}(I_p \otimes i_\tau)]^{-1} [\widehat{B}_j - B_j] \\ &+ \left\{ [(I_p \otimes i_\tau)^\top \widehat{V}^{-1}(I_p \otimes i_\tau)]^{-1} - [(I_p \otimes i_\tau)^\top V^{-1}(I_p \otimes i_\tau)]^{-1} \right\} B_j \\ &+ \left\{ [(I_p \otimes i_\tau)^\top \widehat{V}^{-1}(I_p \otimes i_\tau)]^{-1} - [(I_p \otimes i_\tau)^\top V^{-1}(I_p \otimes i_\tau)]^{-1} \right\} [\widehat{B}_j - B_j].\end{aligned}$$

Therefore, It suffices to prove that $\|\widehat{V}^{-1} - V^{-1}\| = o_p(n^{-\gamma})$, for some $\gamma > 0$. Since the preliminary estimator is \sqrt{n} -consistent we can restrict our attention to the set $\{\theta : \|\theta - \theta_0\| \leq \delta_n \sqrt{n}\}$ for some sequence $\delta_n \rightarrow 0$. It follows that

$$\max_{j,l \in \mathcal{J}_n} \|\widehat{V}_{jl} - V_{jl}\| \leq \left(\max_{j,l \in \mathcal{J}_n} \|\widehat{\Omega}_{jl} - \Omega_{jl}\| + \max_{j \in \mathcal{J}_n} \|\widehat{\Gamma}_j - \Gamma_j\| \right) \left(\min_{j \in \mathcal{J}_n} \lambda_{\min}(\Gamma_j) \right)^{-2} = O_p(n^{-\eta} \tau^{2\rho_1})$$

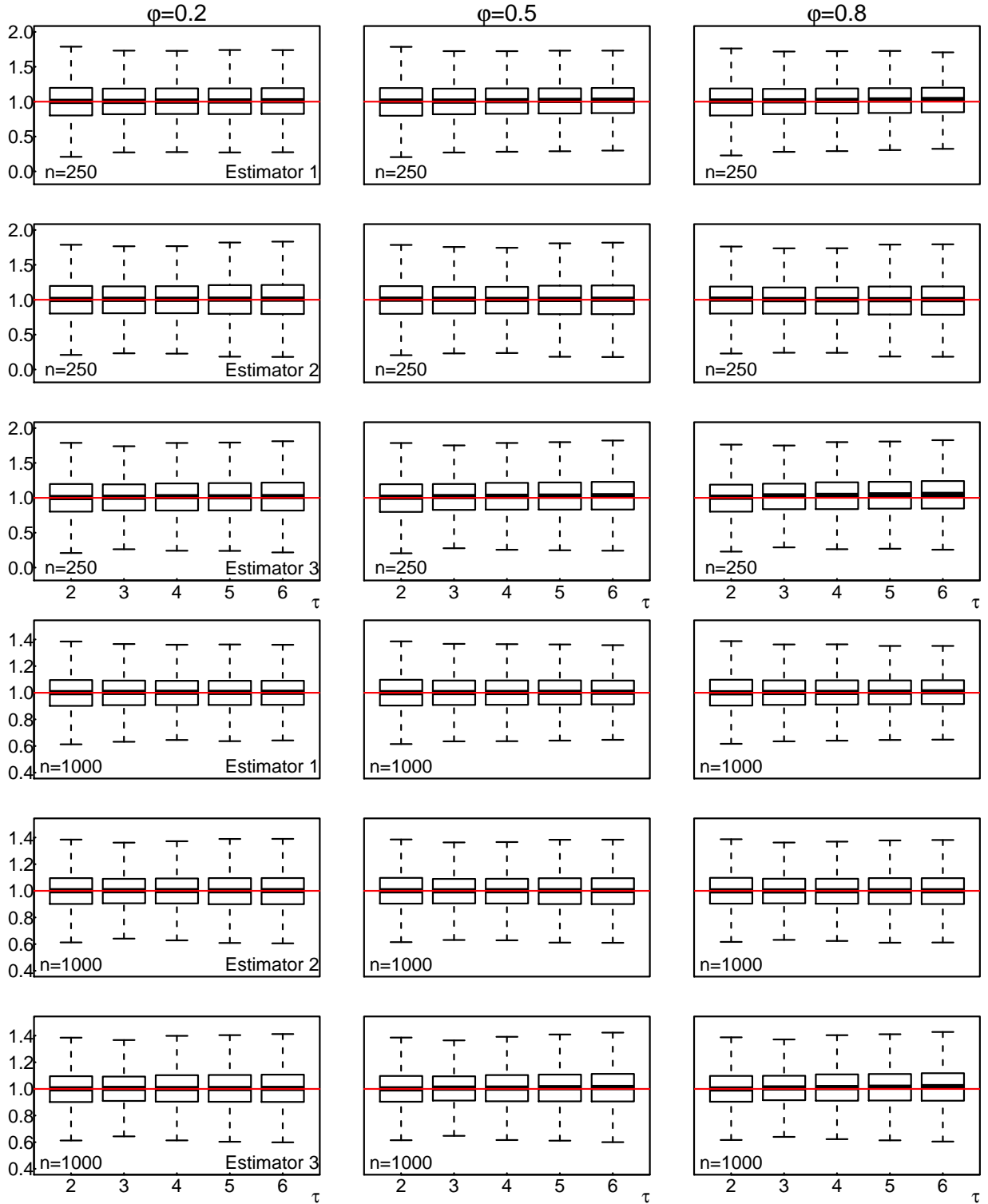
by **D3**.

We have by standard matrix inequalities that

$$\begin{aligned} \|\widehat{V}^{-1} - V^{-1}\| &\leq \lambda_{\max}(\widehat{V}^{-1} - V^{-1}) \\ &\leq \frac{\lambda_{\max}(\widehat{V} - V) \lambda_{\max}(V^{-1})}{1 - \lambda_{\max}(V^{-1}(\widehat{V} - V))} \\ &\leq \tau(n) \max_{j,l \in \mathcal{J}_n} \|\widehat{V}_{jl} - V_{jl}\| \times \frac{1}{\lambda_{\min}(V)} \times \frac{1}{1 - \tau \lambda_{\min}^{-1}(V) \max_{j,l \in \mathcal{J}_n} \|\widehat{V}_{jl} - V_{jl}\|} \\ &= o_p(n^{-(\eta-\epsilon)}), \end{aligned}$$

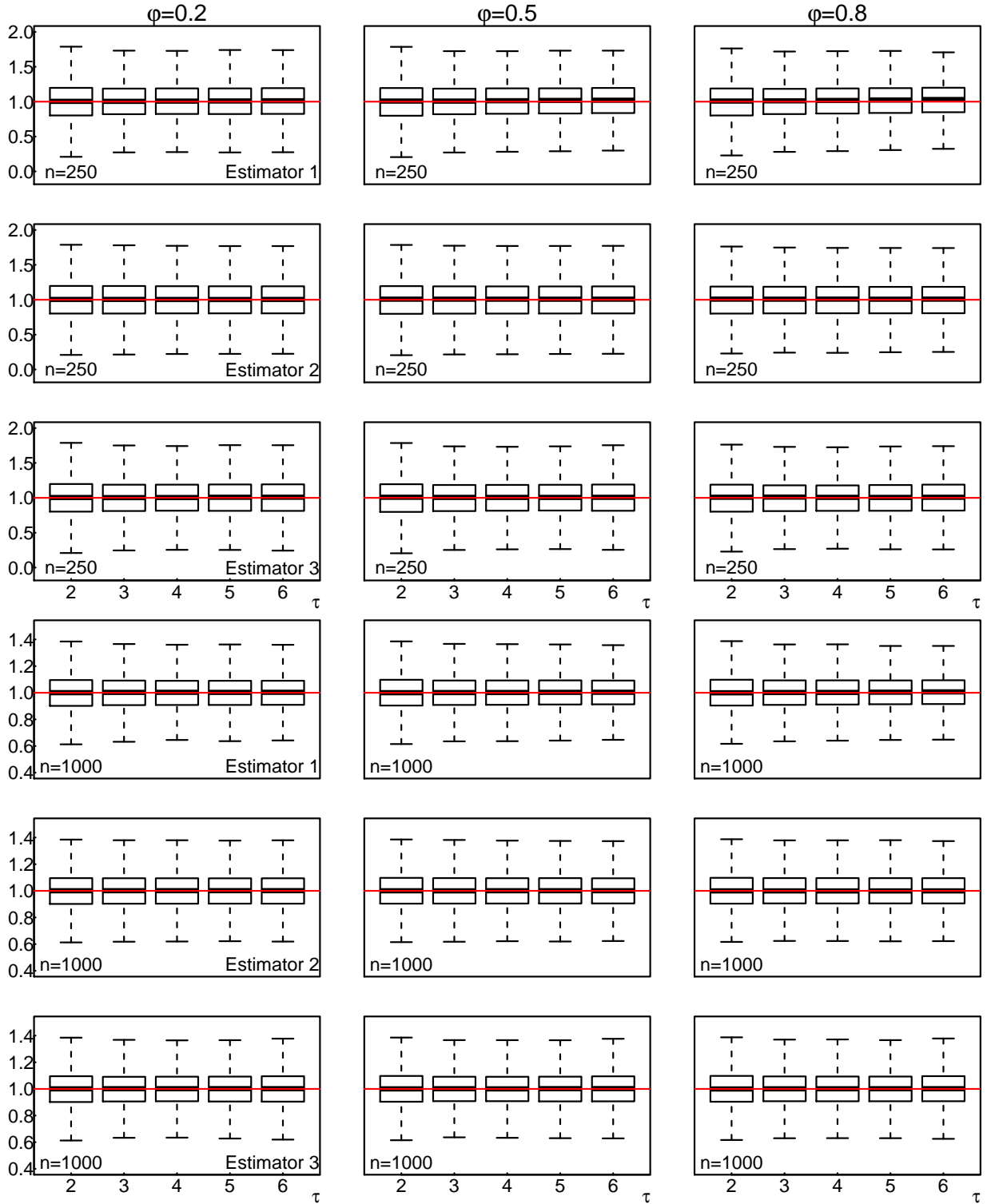
for any strictly positive $\epsilon < \eta$. ■

Figure 1: DGP 1 – Basis 1



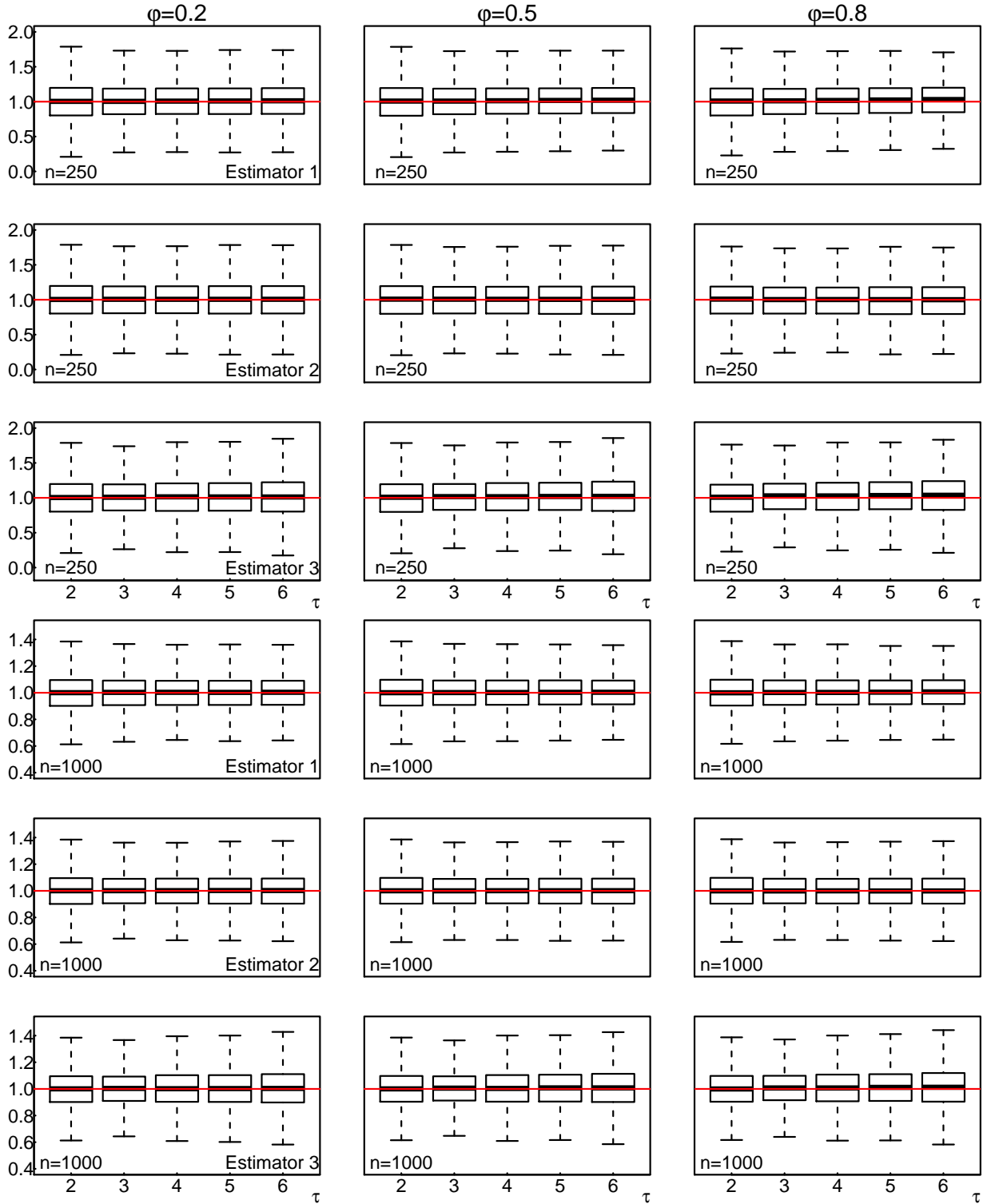
Note: Box plots of 5000 replications of Estimator 1 (Newey, 1990), and the proposed Estimators 2 and 3 with weights $W_j = (j^{-3} / \sum_{j=2}^{\tau} j^{-3}) I_2$ and with weights given by a feasible version of (5.8) respectively for $\tau = 2, \dots, 6$.

Figure 2: DGP 1 – Basis 2



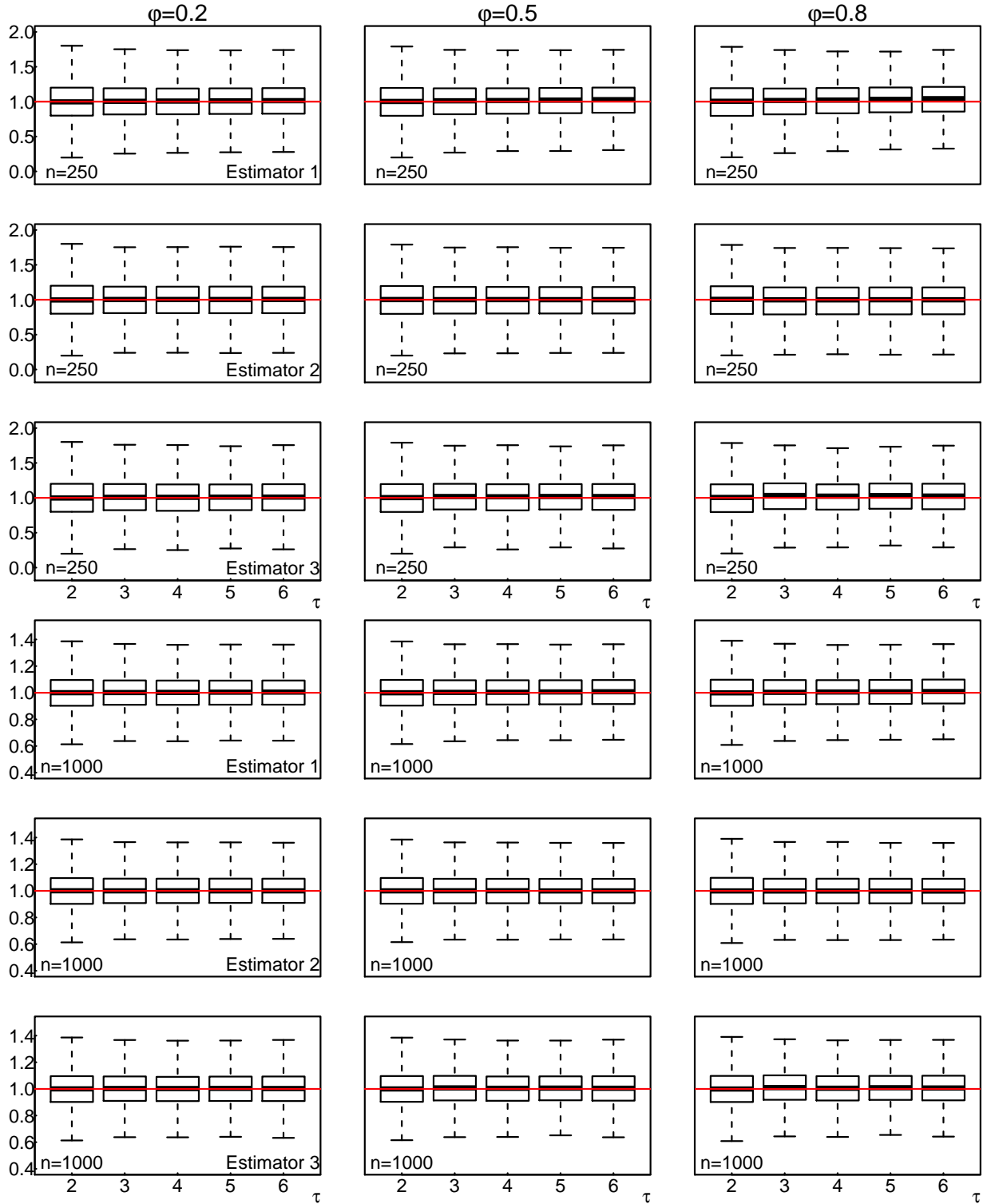
Note: Box plots of 5000 replications of Estimator 1 (Newey, 1990), and the proposed Estimators 2 and 3 with weights $W_j = (j^{-3} / \sum_{j=2}^{\tau} j^{-3}) I_2$ and with weights given by a feasible version of (5.8) respectively for $\tau = 2, \dots, 6$.

Figure 3: DGP 1 – Basis 3



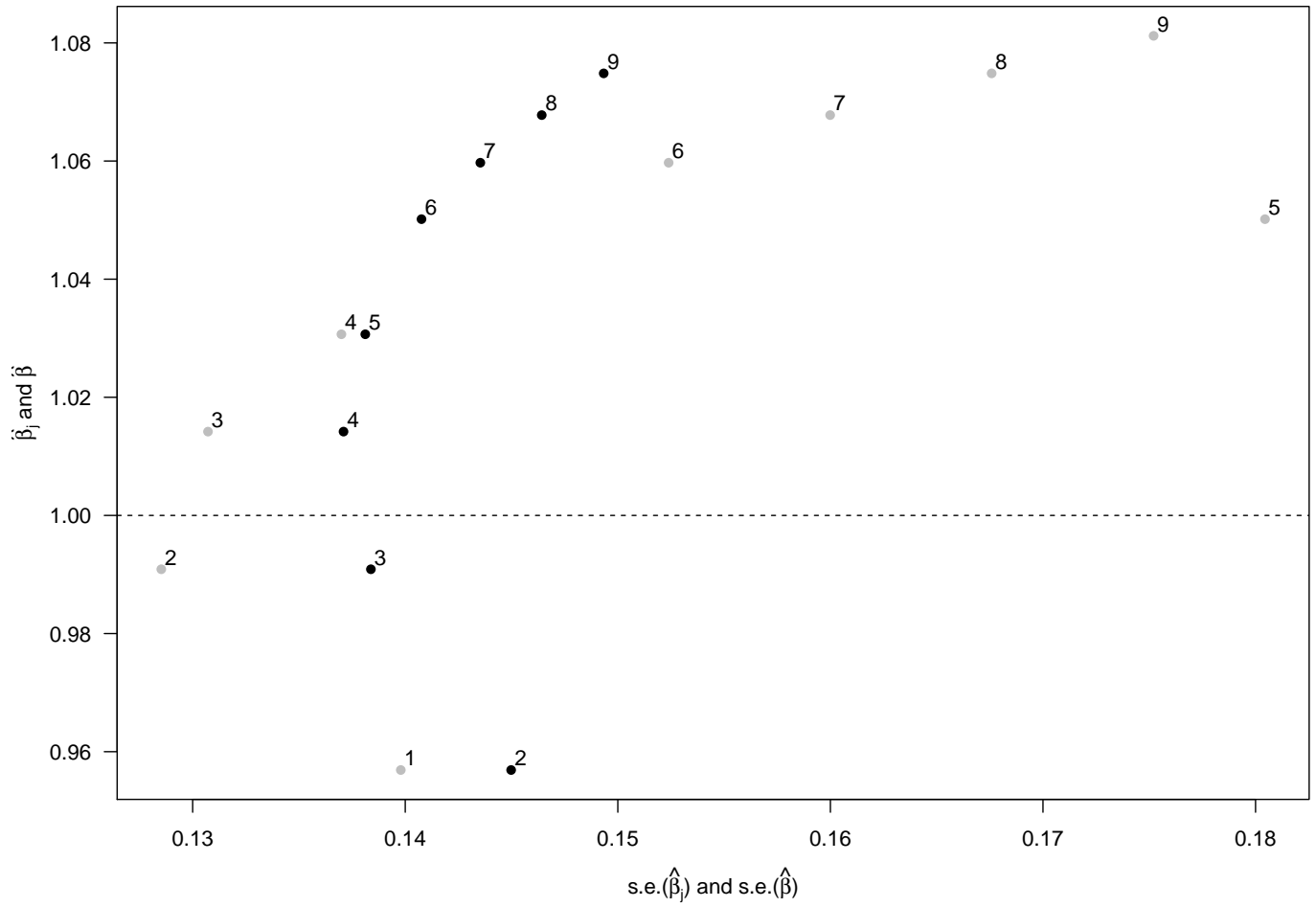
Note: Box plots of 5000 replications of Estimator 1 (Newey, 1990), and the proposed Estimators 2 and 3 with weights $W_j = (j^{-3} / \sum_{j=2}^{\tau} j^{-3}) I_2$ and with weights given by a feasible version of (5.8) respectively for $\tau = 2, \dots, 6$.

Figure 4: DGP1 – Basis 4



Note: Box plots of 5000 replications of Estimator 1 (Newey, 1990), and the proposed Estimators 2 and 3 with weights $W_j = (j^{-3} / \sum_{j=2}^{\tau} j^{-3}) I_2$ and with weights given by a feasible version of (5.8) respectively for $\tau = 2, \dots, 6$.

Figure 5: A Monte Carlo Realization for DGP 1



Note: Gray dots represent a Monte Carlo realization of estimator (7.3) of $\beta_{20; j}$ for $j = 1, \dots, 9$ using Basis 3 with $\rho = 0.8$ and $n = 1000$ against their standard errors (s.e.). Black dots correspond to the proposed estimator (7.2) of $\beta_{20; 2}$ with weights given by a feasible version of (5.8) for $\tau = 2, \dots, 9$ against their standard errors (s.e.). The dotted line represents the true value of β_{20} .

Table 1: DGP2: High-Dimensional Inference

| n | β_{10} | | | β_{20} | | |
|-----------------|--------------|-----------|---------|--------------|-----------|---------|
| | Bias | Std. Dev. | RMSE | Bias | Std. Dev. | RMSE |
| $\varphi = 0.2$ | | | | | | |
| 15 | 0.009 | 0.272 | 0.272 | 0.008 | 0.056 | 0.057 |
| | -0.013* | 0.287* | 0.288* | 0.007* | 0.052* | 0.053* |
| 25 | 0.002 | 0.211 | 0.211 | 0.004 | 0.042 | 0.042 |
| | -0.012* | 0.207* | 0.207* | 0.006* | 0.037* | 0.037* |
| 50 | 0.000 | 0.146 | 0.146 | 0.004 | 0.029 | 0.029 |
| | -0.006** | 0.145** | 0.145** | 0.005** | 0.026** | 0.026** |
| $\varphi = 0.5$ | | | | | | |
| 15 | -0.003 | 0.264 | 0.264 | 0.013 | 0.055 | 0.057 |
| | -0.010* | 0.276* | 0.276* | 0.016* | 0.051* | 0.054* |
| 25 | -0.002 | 0.204 | 0.204 | 0.011 | 0.043 | 0.044 |
| | -0.027* | 0.212* | 0.213* | 0.015* | 0.038* | 0.040* |
| 50 | -0.006 | 0.144 | 0.144 | 0.007 | 0.030 | 0.031 |
| | -0.016** | 0.144** | 0.145** | 0.009** | 0.027** | 0.028** |
| $\varphi = 0.8$ | | | | | | |
| 15 | -0.016 | 0.276 | 0.276 | 0.020 | 0.054 | 0.058 |
| | -0.028* | 0.269* | 0.270* | 0.026* | 0.053* | 0.059* |
| 25 | -0.016 | 0.205 | 0.205 | 0.018 | 0.042 | 0.046 |
| | -0.022* | 0.192* | 0.193* | 0.027* | 0.039* | 0.047* |
| 50 | -0.012 | 0.144 | 0.145 | 0.015 | 0.029 | 0.032 |
| | -0.013** | 0.144** | 0.145** | 0.016** | 0.027** | 0.031** |

Notes: Number of replications = 5000. (*)=OLS estimator. (**)= 2SLS estimator.